Chapter 7 - Sampling Distributions

1 Introduction

What is statistics? It consist of three major areas:

- Data Collection: sampling plans and experimental designs
- Descriptive Statistics: numerical and graphical summaries of the data collected from a sample
- Inferential Statistics: estimation, confidence intervals and hypothesis testing of parameters of interest

Statistical procedures are part (steps 2-5 below) of the **Scientific Method** first espoused by Sir Francis Bacon (1561-1626), who wrote "to learn the secrets of nature involves collecting data and carrying out experiments." The modern methodology:

1. Observe some phenomenon
2. State a hypothesis explaining the phenomenon
3. Collect data
4. Test: Does the data support the hypothesis?
5. Conclusion. If the test fails, go back to step 2.

If you encounter a "scientific claim" that you disagree with, scrutinize the steps of the scientific method used. "Statistics don't lie, but liars do statistics." - Mark Twain.

What is mathematical statistics?: The study of the theoretical foundation of statistics.

What is a statistic? Let $X_1, X_2, \ldots, X_n$ be a set of observable **random variables** (such as a **random sample** of n **individuals** from a **population** of interest). A **statistic** T is a function

$$T = \mathcal{T}(X_1, X_2, \ldots, X_n)$$

applied to $X_1, X_2, \ldots, X_n$.

POPULATION vs. SAMPLE:

> **Population**: The entire group of individuals (subjects or units), that can be either **existent** or **conceptual**, that we want information about.
>
> **Sample**: A part of the population from which data is collected.

PARAMETER vs. STATISTIC:

Parameter: A numerical value calculated from all individuals in the population.

- Population mean: $\mu = \begin{cases} \sum_x xP(x) & \text{if } x \text{ is discrete} \\ \int_{-\infty}^{\infty} xf(x)dx & \text{if } x \text{ is continuous} \end{cases}$
- Population variance: $\sigma^2 = \begin{cases} \sum_x (x-\mu)^2 P(x) & \text{if } x \text{ is discrete} \\ \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx & \text{if } x \text{ is continuous} \end{cases}$
- Population proportion: p is the true proportion of $1's$ in the population.
- Population median: $\phi_{.5}$ is the (not necessarily unique) value such that $P(X \leq \phi_{.5}) \geq .5$ and $P(X \geq \phi_{.5}) \geq .5$.

Statistic: A numerical value calculated from a sample $X_1, ..., X_n$.

- Sample mean: $\overline{X} = \mathcal{T}(X_1, X_2, \ldots, X_n) = \frac{1}{n}\sum_{i=1}^{n} X_i$
- Sample variance: $S^2 = \mathcal{T}(X_1, X_2, \ldots, X_n) = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Sample proportion: $\hat{p} = \mathcal{T}(X_1, X_2, \ldots, X_n) = \frac{\text{number of } 1's}{n}$ is the proportion of 1's in the sample
- Sample median:

$$\hat{\phi}_{.5} = \mathcal{T}(X_1, X_2, \ldots, X_n) = \begin{cases} \text{the middle value if } n \text{ is odd} \\ \text{the average of the two middle values if } n \text{ is even} \end{cases}.$$

2 Sampling Distributions

The value of a statistic varies from sample to sample. In other words, different samples will result in different values of a statistic. Therefore, a statistic is a random variable with a distribution!

Sampling Distribution: The distribution of statistic values from all possible samples of size n. Brute force way to construct a sampling distribution:

- Take all possible samples of size n from the population.
- Compute the value of the statistic for each sample.
- Display the distribution of statistic values as a table, graph, or equation.

2.1 Sampling Distribution of $\overline{X}$

One common population parameter of interest is the population mean μ. In inferential statistics, it is common to use the statistic $\overline{X}$ to estimate μ. Thus, the sampling distribution of $\overline{X}$ is of interest.

Mean and Variance

For any sample size n and a SRS $X_1, X_2, ..., X_n$ from any population distribution with mean μ_x and variance σ_x^2:

- $\mathrm{E}(\overline{X}) = \mu_{\bar{x}} = \mu_x$ and $E(\sum_{i=1}^{n} X_i) = n\mu_x$
- $\mathrm{Var}(\overline{X}) = \sigma^2_{\bar{x}} = \sigma^2_x/n$ and $\mathrm{Var}(\sum_{i=1}^{n} X_i) = n\sigma^2_x$

This result was proved in **Example 5.27** using **Theorem 5.12**: Let a_i for $i = 1, 2, \ldots, k$ be constants and let X_i for $i = 1, 2, \ldots, k$ be random variables. Then

- $\mathrm{E}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i \mathrm{E}(X_i)$ (independence not required) and
- $\mathrm{Var}\left(\sum_{i=1}^{k} a_i X_i\right) = \sum_{i=1}^{k} a_i^2 \,\mathrm{Var}(X_i)$ if $X_1, X_2, \ldots, X_k$ are mutually independent.

Sampling Distribution when the data are normal

For any sample size n and a SRS $X_1, X_2, ..., X_n$ from a normal population distribution $N(\mu_x, \sigma^2_x)$ (**Theorem 7.1**):

- $\overline{X} \sim \mathrm{N}(\mu_x, \sigma^2_x/n)$
- $\sum_{i=1}^{n} X_i \sim N(n\mu_x, n\sigma^2_x)$

Examples:

Suppose that adult male cholesterol levels are distributed as $N(210\text{mg/dL}, (37\text{mg/dL})^2)$.

1. Give an interval centered at the mean μ which captures the middle 95% of all cholesterol values.

2. Give the sampling distribution of $\overline{X}$, the sample mean of cholesterol values taken from SRSs of size $n = 10$.

3. Give an interval centered at the mean μ which captures the middle 95% of all sample mean cholesterol values taken from SRSs of size $n = 10$.

Sampling Distribution for large sample sizes

For a LARGE sample size n and a SRS $X_1, X_2, ..., X_n$ from any population distribution with mean μ_x and variance $\sigma^2_x < \infty$, the approximate sampling distributions are:

$$\overline{X} \,\dot{\sim}\, \mathrm{N}\left(\mu_x, \frac{\sigma^2_x}{n}\right) \text{ and } \sum_{i=1}^{n} X_i \,\dot{\sim}\, \mathrm{N}\left(n\mu_x, n\sigma^2_x\right).$$

This last result follows from the celebrated **Central Limit Theorem**, stated in your book as **Theorem 7.4**:

Let $X_1, X_2, \ldots, X_n$ be a SRS from a distribution with mean μ_x and variance $\sigma_x^2 < \infty$. Then the distribution of

$$U_n = \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{n}}$$

converges to $\mathrm{N}(0,1)$ as $n \to \infty$.

We will prove this theorem later.

Important Examples:

1. **Bernoulli trials**.

   Let $X = \begin{cases} 1 \text{ if } \_\_\_\_\_ \text{ with probability } \quad p = \_\_\_\_\_ \\ 0 \text{ if } \_\_\_\_\_ \text{ with probability} (1-p) = \_\_\_\_\_ \end{cases}$

   Then

   $$X \sim \mathrm{Bern}(p) = \mathrm{Bin}(n = 1, p).$$

   (a) Draw a picture of the pdf of X.

   (b) Find $E(X)$ and $Var(X)$.

   (c) Suppose a SRS X_1, X_2, ..., X_{40} was collected. Give the approximate sampling distribution of $\overline{X}$ (normally denoted by $\hat{p} = \overline{X}$, which indicates that $\overline{X}$ is a sample proportion).

2. **Normal approximation to the Binomial** (section 7.5)

   In the previous example we considered the rv $X \sim \mathrm{Bern}(p) = \mathrm{Bin}(n = 1, p)$. Suppose that a SRS $X_1, X_2, \ldots, X_n$ has been collected with $n > 1$.

   (a) Give the distribution of $Y = \sum_i X_i$, so that Y is the number of successes out of n trials (which is a discrete distribution you learned about in chapter 3).

   (b) Draw a picture of the pdf of $Y = \sum_i X_i$.

(c) Give $E(Y)$ and $Var(Y)$.

(d) In the Example #1c the Central Limit Theorem showed that for any sample size n, when $X \sim \text{Bern}(p)$, then

$$\hat{p} = \overline{X} \dot{\sim} N(\_\_\_\_\_, \_\_\_\_\_).$$

(e) In addition to means $\overline{X}$, the Central Limit Theorem also gives the approximate sampling distribution of a sum $\sum X_i$. Use the Central Limit Theorem to give the approximate sampling distribution of $Y = \sum_i X_i$.

(f) If the true proportion of supporters of healthcare reform in the Montana population is $p = .53$, then out of a SRS of Montanans of size $n = 1000$, whats the probability that less then 500 will pledge support?

2.2 Sampling Distribution of S^2

One common population parameter of interest is the population variance σ^2. In inferential statistics, it is common to use the statistic S^2 to estimate σ^2. Thus, the sampling distribution of S^2 is of interest.

χ^2 **distribution**: The sum of squares of independent standard normal variables is distributed as a χ^2 random random variable. More formally (**Theorem 7.2**):

- If $Z_1, ..., Z_\nu$ are independent and distributed as $N(0, 1)$, then

$$\sum_{i=1}^{\nu} Z_i^2 \sim \chi^2(\nu).$$

  $\chi^2(\nu)$ is called the **chi-square distribution** with ν degrees of freedom.

- For any sample size n and a SRS $X_1, X_2, ..., X_n$ from a normal distribution $N(\mu_x, \sigma_x^2)$,

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu_x}{\sigma_x} \right)^2 \sim \chi^2(n).$$

- Use Table 6 on p. 850 of the textbook for probability calculations.

Mean and Variance

If $C \sim \chi^2(\nu)$, then

- $E(C) = \nu$
- $\text{Var}(C) = 2\nu$.

For any sample size $n > 1$ and a SRS $X_1, X_2, ..., X_n$ from any population distribution with mean μ_x and variance σ_x^2,

- $E(S^2) = \sigma_x^2$
- $\text{Var}(S^2) = \frac{2\sigma_x^4}{n-1}$.

Sampling distribution when the data are normal

For any sample size $n > 1$ and a SRS $X_1, X_2, ..., X_n$ from a normal distribution $N(\mu_x, \sigma_x^2)$ (**Theorem 7.3**):

$$\frac{(n-1)S^2}{\sigma_x^2} \sim \chi^2(n-1)$$

2.3 Sampling Distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$

In inferential statistics, the **test statistic** $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is often used to determine how many **standard errors** $(s/\sqrt{n})$ the sample mean $\overline{X}$ is from a hypothesized value of μ. Thus, the sampling distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$ is of interest.

t distribution (Definition 7.2): If $Z \sim N(0,1)$, $W \sim \chi^2(\nu)$, and Z and W are independent, then:

$$T = \frac{Z}{\sqrt{W/\nu}} \sim t(\nu).$$

$t(\nu)$ is called the **t distribution** with ν degrees of freedom.

Mean and Variance

If $T \sim t(\nu)$, then

- $\text{E}(T) = 0$ for $\nu > 1$
- $\text{Var}(T) = \frac{\nu}{\nu-2}$ for $\nu > 2$

Sampling distribution when the data are normal

For any sample size $n > 1$ and a SRS $X_1, X_2, ..., X_n$ from $N(\mu, \sigma^2)$, then

- $\overline{X}$ and S^2 are independent

- And now by Theorems 7.1 and 7.3 and Definition 7.2

$$T = \frac{\sqrt{n}(\overline{X} - \mu)/\sigma}{\sqrt{\frac{(n-1)S^2/\sigma^2}{n-1}}} = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t(n-1).$$

Sampling Distribution for large sample sizes

For a LARGE sample size n and a SRS $X_1, X_2, ..., X_n$ from any population distribution with mean μ_x and variance $\sigma_x^2 < \infty$:

$$T = \frac{\overline{X} - \mu_x}{S/\sqrt{n}} \dot{\sim} t(n-1).$$

Some useful facts:

- The pdf of T is given by

$$f(t) \;=\; \frac{\Gamma(\frac{\nu+1}{2})}{\Gamma(\frac{\nu}{2})} \frac{1}{\sqrt{\nu\pi}} \left(1 + \frac{t^2}{\nu}\right)^{-(\nu+1)/2}$$

- The t distributions are symmetric about 0 and is bell-shaped like the normal $\mathrm{N}(0,1)$ distribution but with thicker tails.
- As $\nu \to \infty$, the $t(\nu)$ distribution approaches the standard normal distribution.
- Use Table 5 on page 849 for probability calculations.

Examples:
Suppose that adult male cholesterol levels are distributed as $N(210\text{mg/dL}, \sigma^2)$.

1. Give the sampling distribution of $\frac{\overline{X}-\mu}{S/\sqrt{n}}$, where the statistics $\overline{X}$ and S^2 are calculated from a SRS of size $n = 10$.

2. If $S^2 = 36.5^2$, give an interval centered at the mean μ which captures the middle 95% of all sample mean cholesterol values taken from SRSs of size $n = 10$.

2.4 Sampling Distribution of $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$

In inferential statistics, it is often of interest to compare the variances σ_1^2 and σ_2^2 from two populations, and determine if they are different. Based on two SRSs, one of size n_1 with sample variance S_1^2 and the other of size n_2 with sample variance S_2^2, the statistic $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ is often used. Thus, the sampling distribution of $\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$ is of interest.

F distribution (Definition 7.3): If $W_1 \sim \chi^2(\nu_1)$ and $W_2 \sim \chi^2(\nu_2)$ are independent, then:

$$F = \frac{W_1/\nu_1}{W_2/\nu_2} \sim F(\nu_1, \nu_2).$$

$F(\nu_1, \nu_2)$ is called the **F distribution** with ν_1 **numerator degrees of freedom** and ν_2 **denominator degrees of freedom**.

Mean and Variance

If $F \sim F(\nu_1, \nu_2)$, then

- $\mathrm{E}(F) = \frac{\nu_2}{\nu_2 - 2}$ for $\nu_2 > 2$
- $\mathrm{Var}(F) = \frac{2\nu_2^2(\nu_1+\nu_2-2)}{\nu_1(\nu_2-2)^2(\nu_2-4)}$ for $\nu_2 > 4$

Sampling Distribution when the data are normal

If $X_1, X_2, ..., X_{n_1}$ are a SRS from $N(\mu_1, \sigma_1^2)$ and if $Y_1, X_2, ..., X_{n_2}$ are an independent SRS from $N(\mu_2, \sigma_2^2)$, then

$$W_1 = \frac{(n_1-1)S_1^2}{\sigma_1^2} \sim \chi^2(n_1 - 1) \quad \text{and } W_2 = \frac{(n_2-1)S_2^2}{\sigma_2^2} \sim \chi^2(n_2 - 1)$$

are independent, and so

$$F = \frac{W_1/(n_1-1)}{W_2/(n_2-1)} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F(n_1 - 1, n_2 - 1).$$

Some useful facts:

- If $X \sim F(\nu_1, \nu_2)$, then the pdf of X is

$$f(x) = \frac{\Gamma\left(\frac{\nu_1+\nu_2}{2}\right)}{\Gamma\left(\frac{\nu_1}{2}\right)\Gamma\left(\frac{\nu_2}{2}\right)} \left(\frac{\nu_1}{\nu_2}\right)^{\nu_1/2} x^{(\nu_1/2)-1} \left(1 + \frac{\nu_1}{\nu_2}x\right)^{-(\nu_1+\nu_2)/2}.$$

- Use Table 7 on p. 852 for probability calculations.

Example:

Is the variance of female reaction times different than the variance of male reaction times? Jason Paulak at the University of Cincinnati ran a web based reaction experiment to answer this question. $n_1 = 398$ females participated, with $\overline{X}_1 = 517$ ms and $S_1 = 899$ ms. $n_2 = 469$ males participated, with $\overline{X}_2 = 383.2$ ms and $S_2 = 335.7$ ms.

- Give the sampling distribution of $F = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2}$.

- If the two population variances are indeed the same, $\sigma_1^2 = \sigma_2^2$, then what is the probability of observing a ratio of sample variances that we did, or larger?

3 Proof of the Central Limit Theorem

Central Limit Theorem
Let $X_1, X_2, \ldots, X_n$ be a SRS from a distribution with mean μ_x and variance $\sigma_x^2 < \infty$. Then

$$\lim_{n\to\infty} U_n = \lim_{n\to\infty} \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{n}} = U \sim N(0,1).$$

Proof (section 7.4): The proof relies on *moment generating functions* (mgf) from section 3.9 of your textbook. We saw that every rv, as well as having a unique pmf or pdf, also has a unique mgf (**Theorem 7.5**), written as $M(t)$. This notation for the mgf reinforces the fact that it is a function of a real number t in some neighborhood of $t = 0$. The mgf for a continuous rv Y is defined by

$$M(t) = E(e^{tY}) = \int_{-\infty}^{\infty} e^{ty} f(y) dy,$$

where $f(y)$ is the pdf of Y. To prove the Central Limit Theorem:

1. Use Taylor series to find the mgf of $Z_i = \frac{X_i - \mu}{\sigma}$.
2. Find the mgf of $U_n = \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{n}}$.
3. Show that as $n \to \infty$, the mgf of U_n converges to $e^{t^2/2}$.
4. Recall that the mgf of a standard normal rv is $e^{t^2/2}$. Thus, by **Theorem 7.5**, the distribution of $U = \lim_{n\to\infty} U_n$ is a standard normal!

Here we go:

1. Let $Z_i = \frac{X_i - \mu}{\sigma}$. This is where we need the finite variance assumption: if σ^2 is infinite, then Z_i is not well defined!
   (a) Show that $EZ = 0$, $Var(Z) = E(Z^2) = 1$. *Hint*: This is easy since Z is a linear function of X.

(b) Show that $M_Z(0) = 1$, $\frac{d}{dt}M_Z(0) = 0$, and $\frac{d^2}{dt^2}M_Z(0) = 1$.

(c) Use Taylor's Theorem (and the results from (b)) to expand $M_Z(t)$ about $t = 0$ to show that $M_Z(\frac{t}{\sqrt{n}}) = 1 + \frac{(t/\sqrt{n})^2}{2!} + R(\frac{t}{\sqrt{n}})$, where $R(\frac{t}{\sqrt{n}})$ is a remainder of cubic terms of $\frac{t}{\sqrt{n}}$ and higher.

(d) Show that the remainder $\lim_{n\to\infty} \frac{R(\frac{t}{\sqrt{n}})}{(t/\sqrt{n})^2} = n \lim_{n\to\infty} R(\frac{t}{\sqrt{n}}) = 0$.

2. By definition

$$U_n = \frac{\overline{X} - \mu_x}{\sigma_x/\sqrt{n}} = \frac{\sqrt{n}}{n}\left(\frac{\sum_i X_i - n\mu_x}{\sigma_x}\right) = \frac{1}{\sqrt{n}}\sum_i Z_i,$$

which shows that the mgf of U_n when $X_1, ..., X_n$ is a simple random sample is

$$M_{U_n}(t) = \Pi_{i=1}^n M_{Z_i}(\frac{t}{\sqrt{n}}) = \left(M_Z(\frac{t}{\sqrt{n}})\right)^n$$

(by **Theorem 6.2** and Exercise 3.158).

3. Show that $\lim_{n\to\infty} M_{U_n}(t) = e^{\frac{t^2}{2}}$. *Hint:* Substitute in the Taylor series approximation for $M_Z(\frac{t}{\sqrt{n}})$; use the fact that $\lim_{n\to\infty}\left(1 + \frac{a_n}{n}\right)^n = e^{\lim_{n\to\infty} a_n}$; use the fact that $\lim_{n\to\infty} \frac{R(\frac{t}{\sqrt{n}})}{(t/\sqrt{n})^2} = n\lim_{n\to\infty} R(\frac{t}{\sqrt{n}}) = 0$.

4. Recall that if $U \sim N(0,1)$, then $M_U(t) = e^{t^2/2}$ from section 3.9 of your textbook. To prove this, multiply the two exponentials in the integral, then complete the square to show that $E(e^{tU}) = e^{t^2}\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{\infty} e^{-\frac{1}{2}(u-t)^2}du$. Now observe that $\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}(u-t)^2}$ is the pdf of a $N(t,1)$ rv.