Chapter 7

Sampling Distributions & The Central Limit Theorem

Detailed Solutions

7.1 Basic Concept

7.1 Fundamental Definitions (The Big Picture)

Problem: Explain core statistical terms using the workflow diagram.
Solution:

(a) **Probability vs. Statistics:**

- **Probability** moves Left $\rightarrow$ Right. Given the known population parameters (e.g., a fair coin $p = 0.5$), we predict the likelihood of sample outcomes.
- **Statistics** moves Right $\rightarrow$ Left. Given sample data (observed outcomes), we infer the unknown population parameters.

(b) **Parameter vs. Statistic:**

- **Parameter:** A fixed, unknown constant describing the Population (e.g., True mean μ).
- **Statistic:** A random variable calculated from the Sample (e.g., Sample mean $\bar{X}$). It changes with every new sample.

(c) **Population vs. Sample:**

- **Population:** The set of ALL possible observations or items of interest (The Blue Blob).
- **Sample:** A subset of the population selected for analysis (The Green Blob).

(d) **Random Sample vs. Sampling Distribution:**

- **Random Sample:** One specific realization of data points $\{x_1, \ldots, x_n\}$.

- **Sampling Distribution:** The probability distribution of the statistic (e.g., $\bar{X}$) if we were to repeat the sampling process infinitely many times. It describes how the statistic behaves.

(a) Probability: L→R, Statistics: R→L
(b) Parameter: Fixed, Statistic: Random
(c) Population: All, Sample: Subset
(d) Sample: One dataset, Sampling Dist: Distribution of the statistic over many samples.

7.2 The Central Limit Theorem (CLT)

Problem: State and explain CLT.
Solution:

(a) **Statement:** The CLT states that the distribution of the sample mean $\bar{X}$ (or sum) will approximate a Normal distribution as the sample size n becomes large, regardless of the shape of the original population distribution (provided it has a finite variance).

(b) **Shape:** As $n \to \infty$, the sampling distribution of $\bar{X}$ becomes $N(\mu, \sigma^2/n)$.

(c) **Independence:** No. The standard CLT assumes independent and identically distributed (i.i.d.) variables. If there is strong dependence (correlation) between samples, the theorem may not hold or requires modification.

(d) **Simulation vs. Proof:** Mathematical proof involves Characteristic Functions (Taylor series of e^{itx}), which is abstract. Simulation shows the histogram literally "morphing" into a Bell curve, which is visually intuitive and convincing.

(a) Sample mean → Normal as $n \to \infty$.
(b) Bell-shaped (Normal).
(c) No, independence is required.

7.2 Intermediate

7.3 Consistency of the Sample Mean

Problem: Weak Law of Large Numbers (WLLN).
Solution:

1. As $n \to \infty$, the probability that the sample mean deviates from the true mean by any amount ϵ goes to zero.

$$\lim_{n\to\infty} P(|\bar{X}_n - \mu| > \epsilon) = 0$$

2. **Justification:** This guarantees that if we collect enough data, our estimate $\bar{x}$ will be arbitrarily close to the truth μ with high probability. $\bar{X}$ is a "consistent" estimator.

As $n \to \infty$, the error probability $\to 0$. Consistency justifies using $\bar{x}$ to estimate μ.

7.4 Sample Size Determination (Chebyshev)

Problem: $\sigma^2 = 100, \epsilon = 2$, Prob ≥ 0.95.
Solution:

Chebyshev's inequality for sample mean:

$$P(|\bar{X} - \mu| < \epsilon) \geq 1 - \frac{\sigma^2}{n\epsilon^2}$$

We want this probability to be at least 0.95.

$$\begin{aligned} 1 - \frac{100}{n(2^2)} &\geq 0.95 \\ 0.05 &\geq \frac{100}{4n} \\ 0.05 &\geq \frac{25}{n} \\ n &\geq \frac{25}{0.05} = 500 \end{aligned}$$

(Note: If we assumed Normality, n would be much smaller, approx 96. Chebyshev is conservative).

Minimum $n = 500$.

7.5 Sampling Distribution of Mean (Normal Pop)

Problem: $\mu = 4000, \sigma = 200, n = 25$.
Solution:

Since the population is Normal, $\bar{X}$ is exactly Normal $N(\mu, \sigma/\sqrt{n})$.

(a) **Parameters:**

$$E[\bar{X}] = \mu = 4000$$
$$\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{200}{\sqrt{25}} = \frac{200}{5} = 40$$

(b) **Prob $\bar{X} < 3950$:**

$$Z = \frac{3950 - 4000}{40} = \frac{-50}{40} = -1.25$$
$$P(Z < -1.25) \approx 0.1056$$

(c) **Inverse (95% Exceeds):** We want x such that $P(\bar{X} > x) = 0.95 \implies P(\bar{X} < x) = 0.05$. $Z_{0.05} = -1.645$.

$$x = \mu + Z\sigma_{\bar{X}} = 4000 + (-1.645)(40) = 4000 - 65.8 = 3934.2$$

(d) **Sample Size Effect:** Standard Error $SE = \sigma/\sqrt{n}$. To halve SE, we need:

$$\frac{\sigma}{\sqrt{n_{new}}} = \frac{1}{2}\frac{\sigma}{\sqrt{n}} \implies \sqrt{n_{new}} = 2\sqrt{n} \implies n_{new} = 4n$$

$$n_{new} = 4(25) = 100$$

(a) Mean 4000, SE 40
(b) 0.1056
(c) 3934.2 psi
(d) $n = 100$ (Need 4× sample size)

7.6 Sampling Distribution of the Sum

Problem: $\mu = 170, \sigma = 20, n = 16$. Max Load 2800.
Solution:

(a) **Mean and SD of Sum S:**

$$E[S] = n\mu = 16(170) = 2720$$
$$Var(S) = n\sigma^2 \implies \sigma_S = \sqrt{n}\sigma = \sqrt{16}(20) = 4(20) = 80$$

(b) **Distribution:** By CLT, $S \approx N(2720, 80^2)$.

(c) **Prob $S > 2800$:**

$$Z = \frac{2800 - 2720}{80} = \frac{80}{80} = 1.0$$
$$P(Z > 1.0) = 1 - 0.8413 = 0.1587$$

(d) **Assumption Necessity:** No. Since $n = 16$ is reasonably large and we sum independent variables, the CLT ensures the sum is approximately Normal even if individual weights are not (e.g., slightly skewed). However, $n = 16$ is on the border; $n \geq 30$ is the safe rule of thumb, but 16 often suffices for symmetric distributions.

(a) Mean 2720, SD 80
(b) Normal
(c) 0.1587
(d) Not strictly necessary due to CLT.

7.7 Sampling Distribution of Proportion

Problem: $p = 0.08, n = 400$.
Solution:

(a) **Validity:** $np = 400(0.08) = 32$. $n(1-p) = 400(0.92) = 368$. Both are > 10, so Normal approximation is valid. $\hat{P} \sim N(p, \frac{p(1-p)}{n})$.

(b) **Mean and SE:** Mean $\mu_{\hat{p}} = p = 0.08$. SE $\sigma_{\hat{p}} = \sqrt{\frac{0.08(0.92)}{400}} = \sqrt{\frac{0.0736}{400}} = \sqrt{0.000184} \approx 0.01356$.

(c) **Prob $\hat{P} > 0.10$:**

$$Z = \frac{0.10 - 0.08}{0.01356} = \frac{0.02}{0.01356} \approx 1.47$$
$$P(Z > 1.47) = 1 - 0.9292 = 0.0708$$

(d) **Prob $0.06 < \hat{P} < 0.10$:** $Z_{0.10} = 1.47$. $Z_{0.06} = \frac{0.06-0.08}{0.01356} = -1.47$.

$$P(-1.47 < Z < 1.47) = 0.9292 - 0.0708 = 0.8584$$

(a) Normal Approx Valid
(b) Mean 0.08, SE 0.0136
(c) 0.0708
(d) 0.8584

7.8 Sampling Distribution of Variance

Problem: $\sigma^2 = 0.01, n = 20$.
Solution:

(a) **Statistic:**

$$\chi^2 = \frac{(n-1)S^2}{\sigma^2}$$

Degrees of freedom $df = n - 1 = 19$.

(b) **Prob $S^2 > 0.015$:** Convert to χ^2:

$$\chi^2 > \frac{19(0.015)}{0.01} = 19(1.5) = 28.5$$

From χ^2 table ($df = 19$): $P(\chi^2 > 27.2) = 0.10$, $P(\chi^2 > 30.1) = 0.05$. So probability is approx between 0.05 and 0.10 (approx 0.07).

(c) **Critical Values (95%):** Need $\chi^2_{0.975}$ and $\chi^2_{0.025}$ for $df = 19$. Lower: $\chi^2_{0.975} = 8.907$. Upper: $\chi^2_{0.025} = 32.852$. Convert back to S^2: $S^2 = \frac{\chi^2\sigma^2}{n-1}$. $a = \frac{8.907(0.01)}{19} \approx 0.0047$. $b = \frac{32.852(0.01)}{19} \approx 0.0173$.

(a) χ^2 with $df = 19$
(b) ≈ 0.07
(c) $0.0047 < S^2 < 0.0173$

7.9 Ratio of Variances (F-Distribution)

Problem: $\sigma_A^2 = 10, n_A = 16; \sigma_B^2 = 15, n_B = 21$.
Solution:

(a) **Distribution:** F-distribution with degrees of freedom $d_1 = 16 - 1 = 15$ and $d_2 = 21 - 1 = 20$.

(b) **Expected Value of Ratio:** Since S^2 is an unbiased estimator, $E[S^2] = \sigma^2$. However, $E[S_A^2/S_B^2] \neq \sigma_A^2/\sigma_B^2$ exactly due to Jensen's inequality (expectation of ratio $\neq$ ratio of expectations). For F-dist, mean is $d_2/(d_2 - 2)$. Qualitatively: It should cluster around $\sigma_A^2/\sigma_B^2 = 10/15 = 0.66$.

(c) **Prob $S_B^2 > 2S_A^2$:** Rearrange to match F-statistic definition ($F = \frac{S_A^2/\sigma_A^2}{S_B^2/\sigma_B^2}$). We want $P(S_B^2/S_A^2 > 2)$. Invert to get F form? Let's look at the standard F-ratio where numerator is A ($d_1 = 15$) and denominator is B ($d_2 = 20$). The statistic $F_{stat} = \frac{S_A^2/10}{S_B^2/15} = 1.5\frac{S_A^2}{S_B^2}$. Inequality: $S_B^2 > 2S_A^2 \implies \frac{S_A^2}{S_B^2} < 0.5$. Substitute into F: $F_{stat} = 1.5(\frac{S_A^2}{S_B^2}) < 1.5(0.5) = 0.75$. Find $P(F_{15,20} < 0.75)$. Using table relationship $F_{1-\alpha}(d_1, d_2) = 1/F_\alpha(d_2, d_1)$. $P(F < 0.75)$ is large? No, mean is near 1. 0.75 is to the left. Software result: $P(F_{15,20} < 0.75) \approx 0.26$.

(a) $F_{15,20}$
(c) ≈ 0.26

7.10 Difference of Two Means

Problem: $\mu_A = 80, \sigma_A = 5; \mu_B = 75, \sigma_B = 3; n = 50$.
Solution:

Let $D = \bar{X}_A - \bar{X}_B$. Mean $\mu_D = 80 - 75 = 5$. Variance $\sigma_D^2 = \frac{\sigma_A^2}{n_A} + \frac{\sigma_B^2}{n_B} = \frac{25}{50} + \frac{9}{50} = 0.5 + 0.18 = 0.68$. $SD_D = \sqrt{0.68} \approx 0.8246$.

(a) **Prob Diff ≥ 7:**

$$Z = \frac{7 - 5}{0.8246} = \frac{2}{0.8246} \approx 2.425$$
$$P(Z > 2.425) \approx 0.0076$$

(b) **Prob Fan B $>$ Fan A ($D < 0$):**

$$Z = \frac{0 - 5}{0.8246} = -6.06$$
$$P(Z < -6.06) \approx 0$$

It is virtually impossible for Fan B to beat Fan A with these sample sizes.

(a) 0.0076 (b) ≈ 0

7.3 Application

7.11 Convergence of Sample Mean

Problem: Simulation of Convergence.
Solution:

```python
import numpy as np
import matplotlib.pyplot as plt

# 1. Population: Exponential (Mean = 10)
true_mean = 10
pop_data = np.random.exponential(true_mean, 100000)

# 2. Simulation
sample_sizes = range(1, 2000, 5)
sample_means = []

for n in sample_sizes:
    sample = np.random.choice(pop_data, n)
    sample_means.append(np.mean(sample))

# 3. Plot
plt.figure(figsize=(10, 6))
plt.plot(sample_sizes, sample_means, label='Sample Mean', alpha=0.6)
plt.axhline(y=true_mean, color='r', linestyle='--', label='True Mean')
plt.xlabel('Sample Size (n)')
plt.ylabel('Sample Mean')
plt.title('Convergence of Sample Mean to Population Mean')
plt.legend()
plt.show()
```

Interpretation: As n increases, the oscillation of the sample mean around the red line (True Mean) dampens significantly. The values tighten around 10. This visualizes **Consistency**: The estimator converges in probability to the true parameter as sample size grows.

7.12 Visualizing the CLT

Problem: Uniform to Normal.
Solution:

```python
import numpy as np
import matplotlib.pyplot as plt
from scipy.stats import norm

# 1. Population (Uniform 0 to 10)
pop_data = np.random.uniform(0, 10, 100000)

# 2. Draw Samples (n=30)
n = 30
num_simulations = 5000
means = [np.mean(np.random.choice(pop_data, n)) for _ in range(
    num_simulations)]

# 3. Plot
plt.figure(figsize=(10, 6))
plt.hist(means, bins=50, density=True, alpha=0.6, color='g', label='
    Sample Means')

# Overlay Normal PDF
mu = 5
sigma_xbar = np.sqrt(8.333 / n) # Var(Uniform) = (b-a)^2/12 = 100/12 =
    8.333
x = np.linspace(3, 7, 100)
plt.plot(x, norm.pdf(x, mu, sigma_xbar), 'r-', linewidth=2, label='
    Theoretical Normal')

plt.xlabel('Sample Mean')
plt.title(f'Sampling Distribution of Mean (n={n})')
plt.legend()
plt.show()
```

Even though the original data is "Flat" (Uniform), the histogram of the means forms a perfect "Bell Curve". This is the Central Limit Theorem in action.