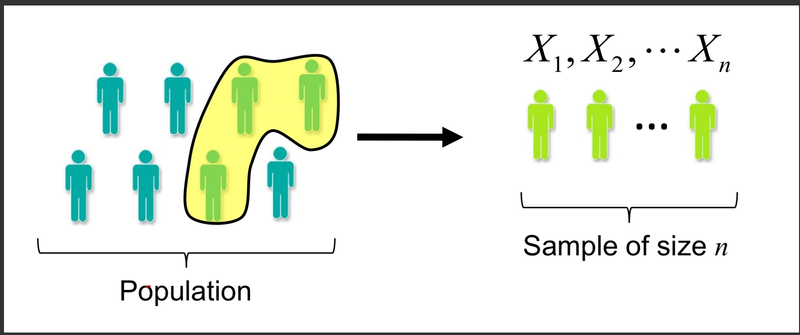
B52 Dec 1 Lec 1 Notes

Statistical Setup

Consider variable of interest (e.g. income) from some population with unknown mean (μ) & variance (σ^2)

We want to estimate mean (μ) without looking at entire population, but using random sampling instead.



Sample Statistics

Statistical analysis relies on probability model for sample data.

Assume sample data, thought of as random quantities rather than values, are i.i.d RVs from "population" distribution

$$X_1, X_2, \cdots X_n \overset{i.i.d.}{\sim} F_X(x)$$

Sample statistics are functions (i.e. transformations) of sample data, related to model parameters.
e.g. Sample mean: $\bar{X}_n = \frac{1}{n}\sum_{n=1}^{n} X_i$

Sampling Distributions

Sample statistics can be used to estimate model parameters.
e.g. we know that $\bar{X}_n \xrightarrow{p} \mu$ as $n \to \infty$ (by WLLN)

Properties of estimation are determined by sampling distribution i.e. distribution of sample statistic
e.g. for sample mean, we know that

$$\bar{X}_n \sim \text{Normal}\left(\mu, \frac{\sigma^2}{n}\right) \quad \text{(by CLT)}$$

Accuracy of estimation improves with sample size n at rate $1/\sqrt{n}$.

Remark: Even though population distribution is unknown, we often assume sample data follow Normal distribution.
- ↳ Most sampling distributions converge to functions of Normal distribution.
- ↳ Easy to calculate probabilities from Normal distribution since any probability can be reduced to standard Normal(0, 1)

Ex 1:

A bottle filling machine is calibrated at a mean of 500 mL, with SD = 4 mL. Find the probability that a random sample of $n = 25$ bottles from a well-calibrated machine gives a sample mean of 500.5 mL or more.

$\bar{X}_n = \frac{1}{n}(X_1 + \dots + X_n)$, where $X_i \sim$ i.i.d. follow $N(500, 4^2)$

From CLT, $\bar{X}_n \overset{\text{approx.}}{\sim} N(\mu, \frac{\sigma^2}{n}) = N(500, 4^2/25 = .8^2)$

$$P(\bar{X}_n > 500.5) = P\left(\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} > \frac{500.5 - 500}{.8}\right) = P(Z > \frac{.5}{.8})$$
$$= 1 - P(Z \le 5/8)$$
$$= 1 - .734 \approx .265986$$

Sample Variance

When invoking CLT, accuracy of $\bar{X}_n$ relies on variance.

For real applications, population variance is generally unknown and must also be estimated.

We can estimate σ^2 by sample variance.

$$S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$$

Proof: Sample variance is unbiased ($E(S_n^2) = \sigma^2$)

$$E(S_n^2) = E\left[\frac{1}{n-1}\sum_{i=1}^{n}(X_i - X_n)^2\right]$$

Since $X_i \overset{\text{i.i.d}}{\sim} N(\mu, \sigma^2)$, $\bar{X}_n \sim N(\mu, \sigma^2/n)$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}((X_i - \mu) - (\bar{X}_n - \mu))^2\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}\left((X_i - \mu)^2 - 2(X_i - \mu)(\bar{X}_n - \mu) + (\bar{X} - \mu)^2\right)\right]$$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2(X_i - \mu)\underbrace{\sum_{i=1}^{n}(X_i - \mu)}_{n(\bar{X} - \mu)} + n(\bar{X} - \mu)^2\right]$$

Since $\bar{X} = \frac{1}{n}\sum_{i=1}^{n}X_i \Rightarrow (\bar{X} - \mu) = \frac{1}{n}\left(\sum_{i=1}^{n}X_i\right) - \mu$

$= \frac{1}{n}\sum_{i=1}^{n}(X_i - \mu)$

$$= \frac{1}{n-1}E\left[\sum_{i=1}^{n}(X_i - \mu)^2 - 2n(\bar{X} - \mu)^2 + n(\bar{X} - \mu)^2\right]$$

$$= \frac{1}{n-1}\left(\sum_{i=1}^{n}\underbrace{E[(X_i - \mu)^2]}_{\sigma^2} - n\underbrace{E[(\bar{X} - \mu)^2]}_{\sigma^2/n}\right)$$

$$= \frac{1}{n-1}\left(n\sigma^2 - n \cdot \frac{\sigma^2}{n}\right) = \sigma^2 \quad \blacksquare$$

Chi - Square Distribution

Let $Z_1, \dots, Z_n \overset{\text{i.i.d}}{\sim} N(0,1)$; then $\begin{cases} Z_1^2 \sim \text{Gamma}(\frac{1}{2}, \frac{1}{2}) = \chi^2(1) \\ Z_1^2 + \dots + Z_n^2 \sim \text{Gamma}(\frac{n}{2}, \frac{1}{2}) = \chi^2(n) \end{cases}$

$\text{Gamma}(\frac{n}{2}, \frac{1}{2})$ is called the Chi-square distribution with parameter n, a.k.a degrees of freedom.

Sampling distribution of sample variance given by

$$\frac{(n-1)S_n^2}{\sigma^2} = \frac{\sum_{i=1}^{n} (X_i - \bar{X}_n)^2}{\sigma^2} \sim \chi^2(n-1)$$

Ex 2:

A bottle filling machine is calibrated at a mean of 500 mL, with SD = 4mL. Find the probability that a random sample of $n=25$ bottles from a well-calibrated machine gives a sample SD of 6 mL or more.

$$P(S_n^2 > 6^2) = P\Big(\underbrace{\tfrac{n-1}{\sigma^2} S_n^2}_{\sim \chi^2(24)} > \tfrac{24}{4^2} \cdot 6^2\Big) = P(\chi^2(24) > 54)$$

$$= .000426243$$

We have to use calculator / software here since χ^2 has no closed form solution