Statistics and Sampling distributions

a **statistic** is a numerical summary of sample data. It is a rv.
The distribution of a statistic is called its **sampling distribution**.

The rv's $X_1, X_2, \cdots, X_n$ are said to form a **random sample** of size n if the $X_i's$ are independent and each X_i has the same probability distribution (identically distributed).

Derive sampling distribution

Example: $P(X=1)=\frac{1}{3}, P(X=0)=\frac{2}{3}$.
Let $\bar{X}=\frac{1}{2}(X_1+X_2)$.
$P(\bar{X}=0)=P(X_1=0,X_2=0)=4/9$.
$P(\bar{X}=\frac{1}{2})=P(X_1=0,X_2=1)+P(X_1=1,X_2=0)=\frac{4}{9}$.
$P(\bar{X}=1)=P(X_1=1,X_2=1)=\frac{1}{9}$.

$\bar{x}$	0	$\frac{1}{2}$	1
$p(\bar{x})$	$\frac{4}{9}$	$\frac{4}{9}$	$\frac{1}{9}$

If $\bar{X}=\frac{1}{3}(X_1+X_2+X_3)$, then

$\bar{x}$	0	$\frac{1}{3}$	$\frac{2}{3}$	1
$p(\bar{x})$	$\frac{8}{27}$	$\frac{4}{9}$	$\frac{2}{9}$	$\frac{1}{27}$

Note $E(\bar{X})=\mu_{\bar{X}}=\frac{1}{3}=E(X)$.

Simulation

Specify

1. The statistic of interest.
2. The population distribution.
3. The sample size n.
4. The number of replication k.

For each sample, compute the statistic and the histogram of the k values gives the approximate distribution of the statistic.

the distribution of the sample mean

proposition: let $X_1, X_2, \cdots, X_n$ be a random sample from a distribution with mean μ and standard deviation σ, then

1. $E(\bar{X}) = \mu_{\bar{X}} = \mu$,
2. $V(\bar{X}) = \sigma^2_{\bar{X}} = \frac{\sigma^2}{n}$ and $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$

Normal population distribution

proposition: Let $X_1, X_2, \cdots, X_n$, iid $\sim N(\mu, \sigma^2)$, then $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$.

the Central Limit Theorem (CLT):
Let $X_1, X_2, \cdots, X_n$ be a random sample from a population with mean μ and variance σ^2. Then as $n \to \infty$,
$\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \to N(0, 1)$.
Roughly speaking, when n is large (≥ 30), $\bar{X}$ approximately $\sim N(\mu, \frac{\sigma^2}{n})$.

The Law of Large Numbers

$\bar{X} \to \mu$ as $n \to \infty$.
example: Flip a coin n times. Let $X_i = 1$ for a head and $X_i = 0$ for a tail. Then $\bar{X}$ = fraction of heads. $\bar{X} \to 0.5$ as $n \to \infty$.

The distribution of a linear combination

Given a collection of n rvs and n constants $a_1, \cdots, a_n$, the rv $Y = a_1X_1 + \cdots + a_nX_n$ is called a **linear combination** of the $X_i's$.

proposition:

1. $E(a_1X_1 + \cdots + a_nX_n) = a_1E(X_1) + a_2E(X_2) + a_nE(X_n)$
2. $V(a_1X_1 + \cdots + a_nX_n) = a_1^2V(X_1) + a_2^2V(X_2) + a_n^2V(X_n)$ if $X_i's$ are independent.

In particular,
$E(X_1 - X_2) = E(X_1) - E(X_2)$ and
$V(X_1 - X_2) = V(X_1) + V(X_2)$ if X_1 and X_2 are independent.

Normal rvs

proposition: If $X_1, X_2, \cdots, X_n$ are independent, normally distributed rv's, then any linear combination of the $X_i's$ also has a normal distribution.

Distributions based on a normal sample

propositions

1. $Z^2 \sim \chi_1^2$.
2. If $X_1 \sim \chi_{\nu_1}^2, X_2 \sim \chi_{\nu_2}^2$, and they are independent, then $X_1 + X_2 \sim \chi_{\nu_1+\nu_2}^2$.
3. $\sum_{i=1}^n Z_i^2 \sim \chi_n^2$ for independent $Z_i' s$.
4. For a random sample from a normal distribution, $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$.
5. If $Z \sim N(0,1), X \sim \chi_\nu^2$, and Z and X are independent, then $T = \frac{Z}{\sqrt{X/\nu}}$ is said to have a t distribution with ν degrees of freedom.

Theorem: If $X_1, X_2, \cdots, X_n$ is random sample from $N(\mu, \sigma^2)$, then $T = \frac{\bar{X}-\mu}{S/\sqrt{n}} \sim t_{n-1}$.

F distribution

F distribution: If $X_1 \sim \chi^2_{\nu_1}, X_2 \sim \chi^2_{\nu_2}$ and X_1 and X_2 are independent, then
$F = \frac{X_1/\nu_1}{X_2/\nu_2} \sim F_{\nu_1,\nu_2}$.
If we have a random sample of size m from $N(\mu_1, \sigma_1^2)$ and an independent sample of size n from $N(\mu_2, \sigma_2^2)$, then
$F = \frac{\frac{(m-1)S_1^2/\sigma_1^2}{m-1}}{\frac{(n-1)S_2^2/\sigma_2^2}{n-1}} = \frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} \sim F_{m-1,n-1}$
F distribution can be used to make inference about the ratio of two population variances.