Lecture 6: Summary

Let X take values $x_1, \ldots, x_n$ with probability $p_i = \text{pr}(X = x_i)$,

$$\begin{aligned} \text{E}X &= \sum x_i p_i \qquad (= \mu_x) \\ \text{Var}(X) &= \text{E}[X - \text{E}X]^2 \\ &= \text{E}[X^2] - [\text{E}X]^2 \\ \text{SD}(X) &= \sqrt{\text{Var}(X)} \end{aligned}$$

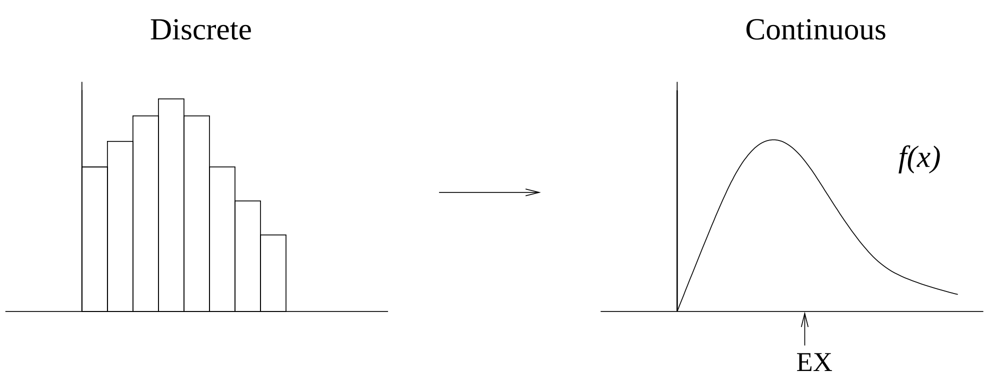
Properties:

1. $\text{E}[\sum_{i=1}^I a_i X_i] = \sum_{i=1}^I a_i \text{E}X_i$
2. $\text{SD}(a + bX) = |b|\ \text{SD}(X)$
3. $\text{Var}[\sum_{i=1}^I a_i X_i] = \sum_{i=1}^I a_i^2 \text{Var}(X_i)$ if $\text{Cov}(X_i, X_j) = 0$ for $i \neq j$

   In general, $\text{Cov}(X_i, X_j) = \text{E}[(X - \text{E}X)(Y - \text{E}Y)]$ which gives,

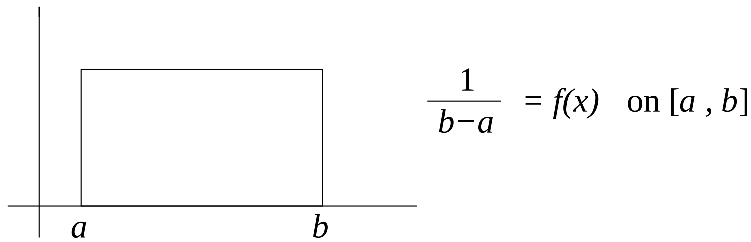
$$\text{Var}(\sum_{i=1}^I a_i X_i) = \sum_{i=1}^I a_i^2 \text{Var}(X_i) + \sum_{i \neq j} a_i a_j \text{Cov}(X_i, X_j)$$

All the properties still hold when X if continuous.



$$\begin{aligned} \text{E}X &= \int_{-\infty}^{\infty} x f(x) dx \qquad = \mu_X \\ \text{Var}(X) &= \text{E}[X^2] - \mu_X^2 \\ &= \int_{-\infty}^{\infty} x^2 f(x) dx - \left[\int_{-\infty}^{\infty} x f(x) dx\right]^2 \end{aligned}$$

e.g. $X \sim$ Uniform on $[a, b]$



$$\mathrm{E}X = \frac{a+b}{2}, \qquad \mathrm{Var}(X) = \frac{(b-a)^2}{12}, \qquad \mathrm{SD}(X) = \frac{|b-a|}{\sqrt{12}}$$

Theorem: $X_1, \ldots, X_n$ - sampling without replacement

(i) $\mathrm{E}(\bar{X}_n) = \mu$

(ii) $\mathrm{Var}(\bar{X}_n) = \frac{1}{n}\sigma^2 + \frac{n-1}{n}\mathrm{Cov}(X_1, X_2)$

(iii) $\mathrm{Cov}(X_1, X_2) = -\frac{\sigma^2}{N-1}$

Proof: (i) $X_1, \ldots, X_n$ have the same distribution as X_1

$$\mathrm{E}(\bar{X}_n) = \mathrm{E}(X_1) = \frac{1}{N}\sum_{i=1}^{N} x_i = \mu$$

(ii) (X_j, X_k), $j \neq k$, have the same distribution.

$$\begin{aligned} \mathrm{Var}(\bar{X}_n) &= \frac{1}{n^2}\sum \mathrm{Var}(X_i) + \frac{1}{n^2}\sum_{j\neq k} \mathrm{Cov}(X_j, X_k) \\ &= \frac{1}{n}\sigma^2 + \frac{n-1}{n}\mathrm{Cov}(X_1, X_2) \end{aligned}$$

Moreover, $\mathrm{Cov}(X_1, X_2) = \frac{-\sigma^2}{N-1}$. Hence

$$\begin{aligned} \mathrm{E}(\bar{X}_n) &= \mu \\ \mathrm{Var}(\bar{X}_n) &= \frac{1}{n}\sigma^2 \frac{N-n}{N-1} \\ \mathrm{SD}(\bar{X}_n) &= \frac{1}{\sqrt{n}}\sigma\sqrt{\frac{N-n}{N-1}} \end{aligned}$$

$\sqrt{\frac{N-n}{N-1}}$ is known as the **correction factor**.

Need to calculate:

X_1	x_1	x_2	$\ldots$	x_N
$\mathrm{pr}(X_1 = x_i)$	$\frac{1}{N}$	$\frac{1}{N}$	$\ldots$	$\frac{1}{N}$

1. $\mathrm{E}X_1 = \frac{1}{N}\sum x_i = \mu$
2. $\mathrm{Var}(X_1) = \left(\frac{1}{N}\sum x_i^2\right) - \mu^2 = \sigma^2$
3. $\mathrm{Cov}(X_1, X_2)$ Assume x_i are distinct

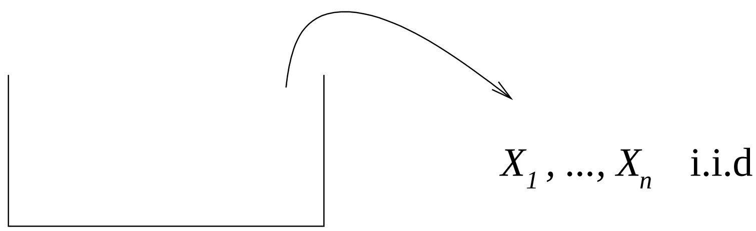
	x_1	x_2	$\ldots$	x_n
x_1	0	$\frac{1}{N}\frac{1}{N-1}$		$\frac{1}{N}\frac{1}{N-1}$
x_2	$\frac{1}{N}\frac{1}{N-1}$	0		
$\vdots$				
$\vdots$				
x_n				0

$$
\begin{aligned}
\mathrm{E}[X_1 X_2] &= \frac{1}{N(N-1)} \sum_{i=1}^{N} x_i \left(\sum_{j=1}^{N} x_j - x_i\right) \\
&= \frac{1}{N(N-1)} \left[\left(\sum_{i=1}^{N} x_i\right)^2 - \sum_{i=1}^{N} x_i^2\right]
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Cov}(X_1, X_2) &= \mathrm{E}[(X_1 - \mathrm{E}X_1)(X_2 - \mathrm{E}X_2)] \\
&= \mathrm{E}[X_1X_2 - X_1\mathrm{E}X_2 - (\mathrm{E}X_1)X_2 + \mathrm{E}X_1\mathrm{E}X_2] \\
&= \mathrm{E}[X_1X_2] - \mathrm{E}X_1\mathrm{E}X_2 - \mathrm{E}X_1\mathrm{E}X_2 + \mathrm{E}X_1\mathrm{E}X_2 \\
&= \mathrm{E}[X_1X_2] - \mathrm{E}X_1\mathrm{E}X_2 \\
&= \frac{1}{N(N-1)} \left[\left(\sum_{i=1}^{N} x_i\right)^2 - \left(\sum_{i=1}^{N} x_i^2\right)\right] - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i\right)^2 \\
&= \left(\frac{1}{N(N-1)} - \frac{1}{N^2}\right)\left(\sum_{i=1}^{N} x_i\right)^2 - \frac{1}{N(N-1)}\sum_{i=1}^{N} x_i^2 \\
&= -\frac{1}{(N-1)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \frac{1}{N^2}\left(\sum_{i=1}^{N} x_i\right)^2\right] \\
&= -\frac{1}{(N-1)}\left[\frac{1}{N}\sum_{i=1}^{N} x_i^2 - \mu^2\right] \\
&= -\frac{\sigma^2}{N-1}
\end{aligned}
$$

$$
\begin{aligned}
\mathrm{Var}(\bar{X}_n) &= \frac{1}{n}\sigma^2 + \frac{n-1}{n}\mathrm{Cov}(X_1, X_2) \\
&= \frac{\sigma^2}{n} + \frac{n-1}{n}\left(-\frac{1}{N-1}\right)\sigma^2 \\
&= \frac{\sigma^2}{n}\left[\frac{N-n}{N-1}\right]
\end{aligned}
$$

"Correction" relative to Sampling with Replacement



$\mathrm{E}\bar{X}_n = \mu$, $\mathrm{Var}(\bar{X}_n)=\frac{\sigma^2}{n}$, when $\frac{n}{N} \to 0$, sampling without replacement $\cong$ sampling with replacement.

Correction factor:

$$
\sqrt{\frac{N-n}{N-1}} \cong \sqrt{1-\frac{n}{N}} \to 1 \quad \text{as} \quad \frac{n}{N} \to 0.
$$

e.g. $N = 1000$, $n = 2$, then $\sqrt{\frac{N-n}{N-1}} = \sqrt{\frac{988}{999}} = 0.9995 = 1$

e.g. $N = 314$, $n = 91$, then $\sqrt{\frac{N-n}{N-1}} = 0.84$.

Report of result: $\bar{X}_n \pm \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}}$

We have:

$$
\mathrm{SD}(\bar{X}_n) = \frac{\sigma}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}
$$

Let σ, the population standard deviation, be unknown.

Estimators for standard errors

Estimate σ^2 and plug-in

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

$E(s^2) = \sigma^2$ if $X_1, \ldots, X_n$ are independent and identically distributed. $X_1, \ldots, X_n$ - sampling without replacement.

$$\begin{aligned} E(s^2) &= \frac{N}{N-1}\sigma^2 \\ \text{so } E\left[s^2 \frac{N-1}{N}\right] &= \sigma^2 \\ E\left[\frac{s^2}{n}\frac{N-n}{N}\right] &= \mathrm{Var}(\bar{X}_n) \end{aligned}$$

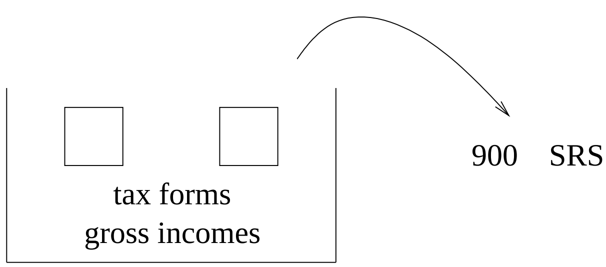
0-1 Population

$$x_1 = \{0,1\} \qquad \pi = \frac{\sum_{i=1}^{N} x_i}{N} \qquad [\sigma^2 = \frac{1}{N}\sum_{i=1}^{N} x_i^2 - (\mu)^2]$$

Now, $\mu = \pi$ and $\sigma^2 = \pi(1-\pi)$.

$$\bar{X}_n \pm \frac{\sqrt{\bar{X}_n(1-\bar{X}_n)}}{\sqrt{n-1}}\sqrt{\frac{N-n}{N}}$$

Example



$N = 50{,}000, \qquad \bar{X}_n = 37.5$ K, $\qquad s = 19$ K

$\mu = ?$: Estimate average gross income with a measure of sampling error:

$$\bar{X}_n \pm \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N-1}}$$

$$37.5 \pm \frac{19}{\sqrt{900}} \times 0.99 = 37.5 \pm 0.627\text{K}$$