Properties of the sample mean

0. Reminder: the Normal distribution

(Appendix B, section B.5)

Aside: a probability density function (pdf) of a continuous random variable, is a function that describes the relative likelihood for this random variable to take on a given value.

The **normal distribution** is a very commonly occurring continuous probability distribution. For example, the distribution of grades on a test administered to many people is normally distributed. It is the most important and the most widely used distribution in statistics.

The pdf of the normal distribution family is given by:

$$f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

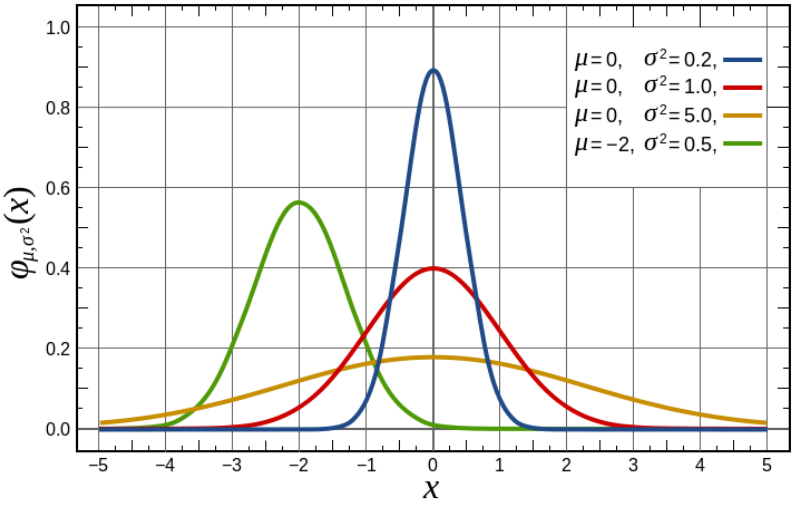
where μ – mean and σ^2 – variance.

The **normal distribution** pdf is a *bell shaped curve*.

Notation:

$X \sim \mathbb{N}(\mu; \sigma^2)$ **(normal distribution)**

$X \sim \mathbb{N}(0; 1)$ **(standard normal distribution)**



Sampling distribution of the sample mean

(Appendix C, section C.2 and C.3)

E.g. tossing a die (uniform distribution)

Population mean E(X) = 3.5

Suppose we take five samples of size *N*=5 and calculate the sample mean.

What does this mean?

Let's toss five dice (this is equivalent to drawing 5 observations from the uniform distribution, i.e. taking a random sample from this distribution). Then let's observe the numbers obtained and compute the sample mean for the first sample. Then do the same for the other 4 samples.

Sample, *N*=5	***Xi***	**Sample mean, $\bar{X}$**
S1 – 1st sample of size 5	4, 1, 1, 6, 6	$\bar{X}$**1**= 3.6 – the sample mean from the 1st sample
S2 – 2nd sample of size 5	5, 4, 1, 6, 5	$\bar{X}$**2** =4.2
S3 – 3rd sample of size 5	6, 1, 2, 1, 2	$\bar{X}$**3** =2.4
S4 – 4th sample of size 5	6, 5, 4, 2, 3	$\bar{X}$**4** =4.0
S5 – 5th sample of size 5	2, 5, 1, 1, 2	$\bar{X}$**5** =2.2

(You can do this with the Excel file uploaded on Canvas).

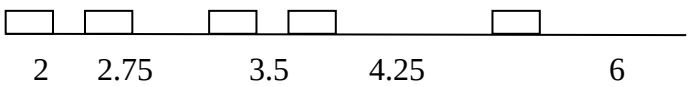
Q: What do we notice?

1) *Is the sample mean the same in all samples?* – Not same in all samples.

2) *Is it always exactly equal to the population mean?* – Not always.

If you compute the mean of many sample, the value of the mean we obtain from each sample will not always be the same and will not always equal the population mean exactly; by chance it will be a little bit higher or a little bit lower.

Now let us draw a frequency *distribution of the sample mean*:



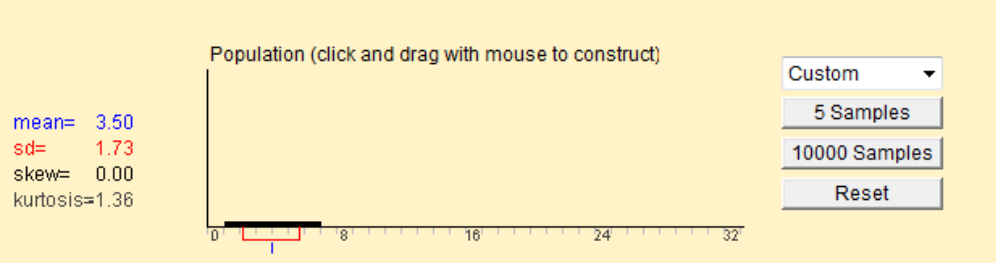
Suppose that we keep doing this and plotting the values on a graph.

KEY: the sample mean $\bar{X}$ itself is random variable (if we compute the mean of many samples, the number we obtain from each sample will generally be a different number**). The sample mean $\bar{X}$ has its own distribution; we call it the sampling distribution of the sample mean.**

The sampling distribution of the mean is a theoretical distribution that is approached as the number of samples approaches infinity.

Online: http://onlinestatbook.com/stat_sim/sampling_dist/index.html

(*Note:* You need Java to run the applet).



Now let's get say 10,000 samples of size 5, calculate the sample mean and keep plotting the mean. Compare to sample of size 25. This means we'll have 10,000 little squares. Each of these dots represents a certain sample mean. You would find that some sample means come much closer to the population mean than others.

What do we notice?

1) The mean of the sampling distribution of the mean is equal to the pop mean → *unbiasedness*

Property 1: Unbiased

Population with mean μ and variance σ^2.

Random sample: $X_1, X_2, \ldots, X_N$.

DEFINITION: Given a population with a mean of μ and variance σ^2, the sampling distribution of the sample mean has an expected value $\mathbf{E(\bar{X}) = \mu}$ – the population mean. **(Unbiased)**

Proof :

$\mathbb{E}(\bar{X}) = \mathbb{E}(\frac{X_1+X_2+\cdots+X_N}{N}) = \frac{1}{N}\mathbb{E}(X_1+X_2+\cdots+X_N) = \frac{1}{N}[\mathbb{E}(X_1)+\mathbb{E}(X_2)+\cdots+\mathbb{E}(X_N)] =$
$\frac{1}{N}(\mu+\mu+\cdots+\mu) = \frac{1}{N}.N\mu = \mu$

Counterexample: biased (median with a skewed population distribution).

2) Compare sample of size 5 to that of size 25: the SD is lower when the sample size is larger → ***efficiency***

Property 2: Efficient

DEFINITION: Efficient – The sample mean $\bar{X}$ has the lowest possible variance among <u>any unbiased</u> estimator of the population mean.

$$\mathbf{Var}(\bar{X}) = \sigma^2/N$$

(or standard deviation $\mathbf{sd}(\bar{X}) = \frac{\sigma}{\sqrt{N}}$).

→ demonstrate with applet (i.e. the spread of the sampling distribution of the mean decreases as the sample size increases).

Proof: Var $(\bar{X}) = \sigma^2/N$

$\mathbf{var}[\frac{1}{N}(X_1+X_2+\ldots+X_N)] = (\frac{1}{N})^2[\mathbf{var}(X_1)+\mathbf{var}(X_2)+\ldots+\mathbf{var}(X_N)] = \frac{1}{N^2}\cdot N\sigma^2 = \frac{\sigma^2}{N}.$

E.g. $(X_1+X_2)/2$ is an unbiased estimator but it *not* efficient. (Every estimator which calculates an average is unbiased).

Property 3: BLUE

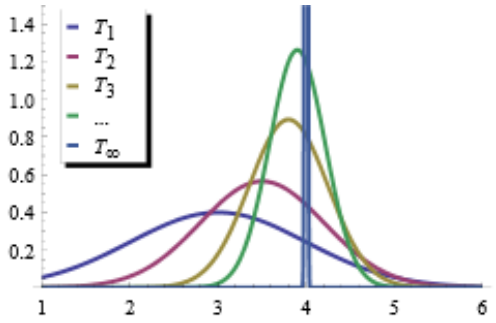
DEFINITION: Best linear unbiased estimator – the sample mean has the smallest variance amongst all *unbiased* estimators of the population mean, which are a linear function of the data.

Properties 1, 2 & 3 are so called ***finite sample properties*** of the sample mean (they refer to a sample with a given sample size *N*).

The next set of properties are so called ***asymptotic properties*** of the sample mean (they explain what happens as the sample size gets infinitely large).

Property 4: Consistent

DEFINITION: as the sample size gets infinitely large (as N → ∞) the sampling distribution of the sample mean collapses to a single point – the true population mean, **μ**.



Applet and STATA: illustrate

Formally: $plim(\bar{X}) = \mu$ (*plim* – probability limit).

This is the same statement as:

- $\bar{X}$ is unbiased.
- $var(\bar{X}) \to 0$ as $N \to \infty$.

This is the so called Law of large numbers (LLN).

What does the LLN mean in simple terms?

➔ **If we are interested to learn about the population mean *μ*, we can get arbitrarily close to it by selecting a sufficiently large sample, and then calculate sample mean.** (Recall the example from class 1). (Demonstrate with Excel).

Why is the LLN important?

➔ **It allows statistical inference when working with samples, rather than entire populations.**

Property 5: (Asymptotic) normality / The Central Limit Theorem (CLT)

Applet and STATA: the sampling distribution on the mean looks like a Normal distribution regardless of what the pop distribution is.

The CLT states that given a population with a mean μ and variance σ², the sampling distribution of the standardized sample mean approaches a standard normal distribution as the sample size increases:

$$\frac{\bar{X}-\mu}{\sigma/\sqrt{N}} \rightarrow \mathbb{N}(0;1)$$

The amazing thing about the central limit theorem is that **no matter what the shape of the original distribution, the sampling distribution of the mean approaches a normal distribution.**

Why is the CLT important?

➔ **It allows testing hypotheses.**

Lastly, these are the general definitions of *unbiasedness* and *consistency* for any estimator $\hat{\mu}$ of a population parameter μ (in our case this is the population mean):

1) An estimator $\hat{\mu}$ of a population parameter μ is ***unbiased*** if $\mathbf{E}(\hat{\mu})=\mu$ for all the possible values of μ (i.e. its expected value equals the true population parameter). Otherwise, we say the estimator is ***biased.*** Its bias is defined as: $\mathbf{Bias}(\hat{\mu})=\mathbf{E}(\hat{\mu})-\mu$.

2) An estimator $\hat{\mu}$ of a population parameter μ is ***consistent*** if $\boldsymbol{plim}(\hat{\mu})=\mu$ for all the possible values of μ (i.e. as the sample size gets infinitely large the value of $\hat{\mu}$ gets arbitrarily close to the true population parameter. This is the same statement as:

- $\text{Bias}(\hat{\mu})\rightarrow 0$ as $N\rightarrow\infty$.
- $\text{var}(\hat{\mu})\rightarrow 0$ as $N\rightarrow\infty$.