The Sampling Distribution of the Mean (SDM).

Take some population X, pick a sample size n, and imagine computing the mean $\bar{x}$ of every size-n sample from X. This gives a new random variable

$$\overline{X} = \{\text{means } \bar{x} \text{ of all size-}n \text{ samples from } X\}.$$

Question: how is $\overline{X}$ distributed?

Answer: intuitively, we'd expect that

(1) The mean $\bar{\mu}$ of $\overline{X}$ equals the mean μ of X ("the average of the averages equals the average");

(2) The std dev $\bar{\sigma}$ of $\overline{X}$ is less than the std dev σ of X (because averaging mitigates the effect of outliers, and therefore reduces spread).

We'd be right! Namely, we have

Theorem (SDM). If X is just about any (not necessarily normal) rv with mean μ and std dev σ, and $n \geq 30$, then the rv

$$\overline{X} = \{\text{means } \bar{x} \text{ of all size-}n \text{ samples from } X\}$$

is roughly $N(\bar{\mu}, \bar{\sigma})$, where

$$\bar{\mu} = \mu \quad \text{and} \quad \bar{\sigma} = \sigma/\sqrt{n}.$$

(Cool fact: this follows from the Central Limit Theorem.)

Example 1.
A set X of 1399 ALEKS Calc readiness exam scores has mean $\mu = 56.81$ and std dev. $\sigma = 22.47$. How is the set $\overline{X}$ of all size-30 sample means from X distributed?

Solution. By SDM, $\overline{X}$ is roughly

$$N(\bar{\mu}, \bar{\sigma}) = N(\mu, \sigma/\sqrt{n}) = N(56.81, 22.47/\sqrt{30})$$

$$= N(56.81, 4.10).$$

NOTE: The number of size-30 sample means from X is

$$\binom{1399}{30} > 6.5 \times 10^{61} \;!!$$

This is HUGE: e.g. the universe, in grams, has mass $< 10^{59}$! So we can't actually compute $\bar{x}$ for every size-30 sample from X. SDM tells us we don't have to!

Example 2 (moving towards statistical inference).

A population X has mean $\mu = 75$ and std. dev. $\sigma = 18$. What's the probability that a size-100 random sample from X has mean between 70 and 80?

Solution.

The question is: what is $P(70 < \overline{X} < 80)$?

We compute:

standardize

$$P(70<\overline{X}<80) = P\left(\frac{70-\mu}{\overline{\sigma}} < \frac{\overline{X}-\mu}{\overline{\sigma}} < \frac{80-\mu}{\overline{\sigma}}\right)$$

use SDM

$$= P\left(\frac{70-75}{18/\sqrt{100}} < \frac{\overline{X}-\overline{\mu}}{\overline{\sigma}} < \frac{80-75}{18/\sqrt{100}}\right)$$

$$= P\left(-2.778 < \frac{\overline{X}-\overline{\mu}}{\overline{\sigma}} < 2.778\right) = 0.995 = 99.5\%.$$

look it up,
use a calculator,
etc.

Example 3.

A population X has unknown mean μ and known std dev. $\sigma = 18$. (It's unlikely that we'd know σ but not μ. But let's pretend, for now.) Suppose a size-100 sample from X has mean $\overline{x} = 83$.

Someone claims the true population mean μ of X is $\mu = 75$. How do you feel about this claim?

Answer: AS IF!!

By Example 2, if the true mean were $\mu = 75$, there would only be a 0.5% chance of a random sample mean $\bar{x}$ being outside the

the interval $(70, 80)$.

So if we <u>do</u> get such an $\bar{x}$, we're pretty confident that $\mu \neq 75$!

This idea is key to statistical inference!