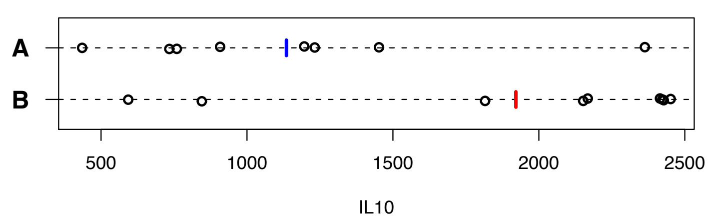
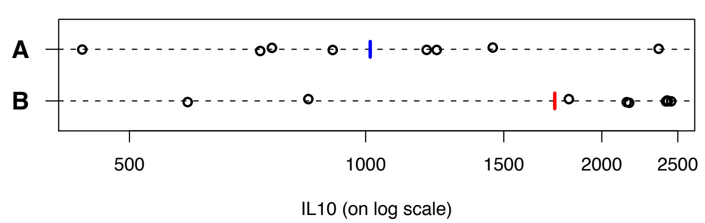
Sampling Distributions

Example

Two strains of mice: A and B.
Measure cytokine IL10 (in males, all same age) after treatment.





⟶ We're not interested in these particular mice, but in aspects of the distributions of IL10 values in the two strains.

Populations and samples

⟶ We are interested in the distribution of measurements in the underlying (possibly hypothetical) population.

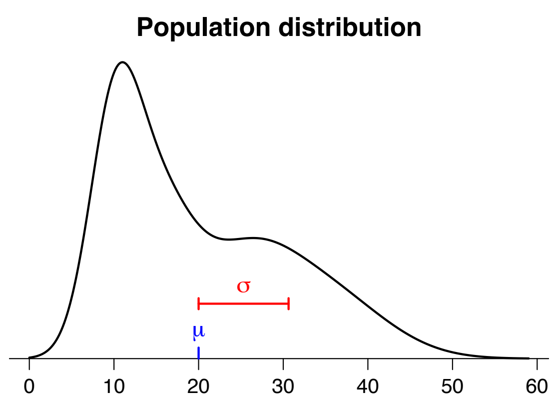
Examples:

- ○ Infinite number of mice from strain A; cytokine response to treatment.
- ○ All T cells in a person; respond or not to an antigen.
- ○ All possible samples from the Baltimore water supply; concentration of cryptospiridium.
- ○ All possible samples of a particular type of cancer tissue; expression of a certain gene.

⟶ We can't see the entire population (whether it is real or hypothetical), but we can see a random sample of the population (perhaps a set of independent, replicated measurements).

Parameters

We are interested in the population distribution or, in particular, certain numerical attributes of the population distribution, called parameters.



⟶ Examples:

- ○ mean
- ○ median
- ○ SD
- ○ proportion = 1
- ○ proportion $>$ 40
- ○ geometric mean
- ○ 95th percentile

Parameters are usually assigned greek letters (like θ, μ, and σ).

Sample data

We make n independent measurements (or draw a random sample of size n). This gives $X_1, X_2, \ldots, X_n$ independent and identically distributed (iid), following the population distribution.

$\longrightarrow$ Statistic:
A numerical summary (function) of the X's. For example, the sample mean, sample SD, etc.

$\longrightarrow$ Estimator:
A statistic, viewed as estimating some population parameter.

We write:

$\overline{X} = \hat{\mu}$ as an estimator of μ, $S = \hat{\sigma}$ as an estimator of σ, $\hat{p}$ as an estimator of p, $\hat{\theta}$ as an estimator of θ, ...

Parameters, estimators, estimates

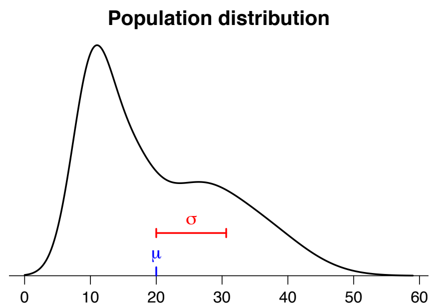
μ
- The population mean
- A parameter
- A fixed quantity
- Unknown, but what we want to know

$\overline{X}$
- The sample mean
- An estimator of μ
- A function of the data (the X's)
- A random quantity

$\overline{x}$
- The observed sample mean
- An estimate of μ
- A particular realization of the estimator, $\overline{X}$
- A fixed quantity, but the result of a random process.

Estimators are random variables

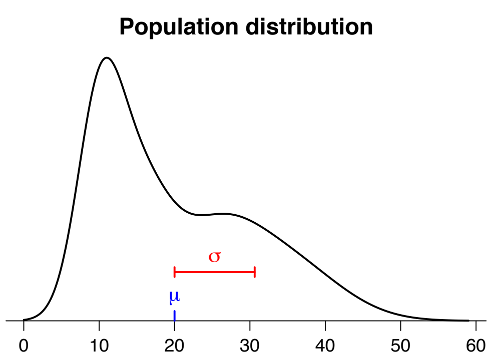
Estimators have distributions, means, SDs, etc.



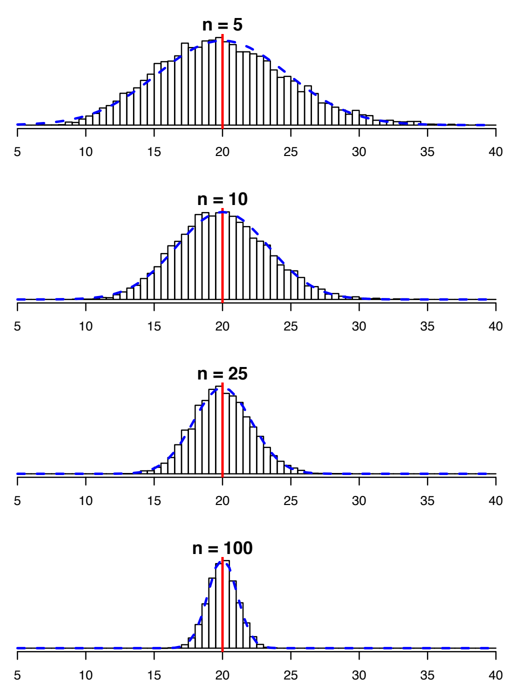
$$\longrightarrow \quad X_1, X_2, \ldots, X_{10} \quad \longrightarrow \quad \overline{X}$$

3.8	8.0	9.9	13.1	15.5	16.6	22.3	25.4	31.0	40.0	$\longrightarrow$	18.6
6.0	10.6	13.8	17.1	20.2	22.5	22.9	28.6	33.1	36.7	$\longrightarrow$	21.2
8.1	9.0	9.5	12.2	13.3	20.5	20.8	30.3	31.6	34.6	$\longrightarrow$	19.0
4.2	10.3	11.0	13.9	16.5	18.2	18.9	20.4	28.4	34.4	$\longrightarrow$	17.6
8.4	15.2	17.1	17.2	21.2	23.0	26.7	28.2	32.8	38.0	$\longrightarrow$	22.8

Sampling distribution



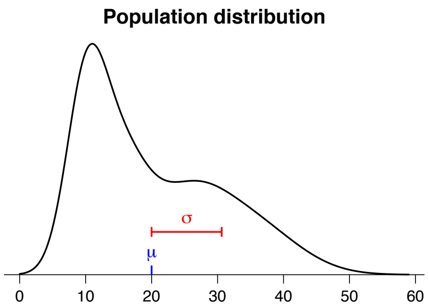
Distribution of $\overline{X}$

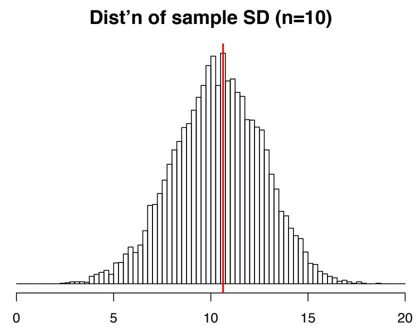


The sampling distribution depends on:

- The type of statistic
- The population distribution
- The sample size

Bias, SE, RMSE





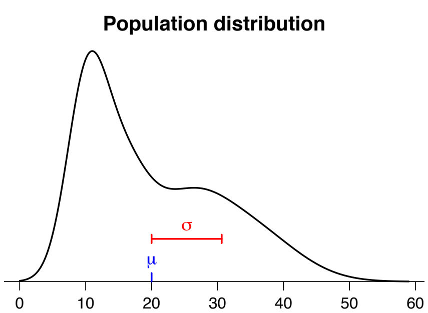
Consider $\hat{\theta}$, an estimator of the parameter θ.

$\longrightarrow$ Bias: $\mathsf{E}(\hat{\theta} - \theta) = \mathsf{E}(\hat{\theta}) - \theta$.

$\longrightarrow$ Standard error (SE): $\mathsf{SE}(\hat{\theta}) = \mathsf{SD}(\hat{\theta})$.

$\longrightarrow$ RMS error (RMSE): $\sqrt{\mathsf{E}\{(\hat{\theta} - \theta)^2\}} = \sqrt{(\mathsf{bias})^2 + (\mathsf{SE})^2}$.

The sample mean



Assume $X_1, X_2, \dots, X_n$ are iid with mean μ and SD σ.

$\longrightarrow$ Mean of $\overline{X}$ = E($\overline{X}$) = μ.

$\longrightarrow$ Bias = E($\overline{X}$) – μ = 0.

$\longrightarrow$ SE of $\overline{X}$ = SD($\overline{X}$) = $\sigma/\sqrt{n}$.

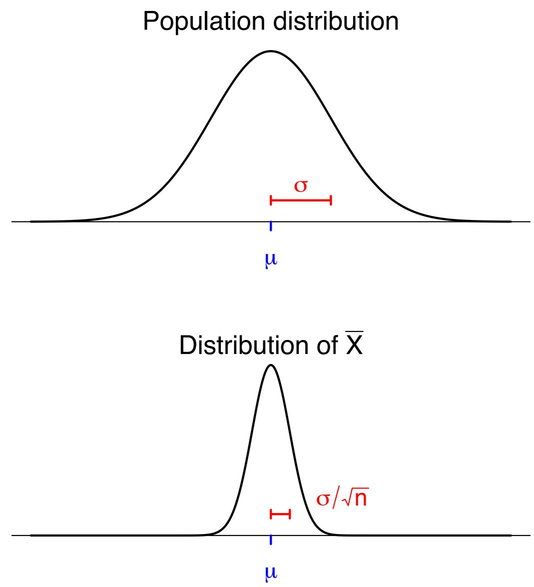
$\longrightarrow$ RMS error of $\overline{X}$:

$\sqrt{(\mathsf{bias})^2 + (\mathsf{SE})^2} = \sigma/\sqrt{n}$.

If the population is normally distributed

If X_1, X_2, ..., X_n are iid Normal(μ,σ), then

$\longrightarrow$ $\overline{X} \sim$ Normal$(\mu, \sigma/\sqrt{n})$.

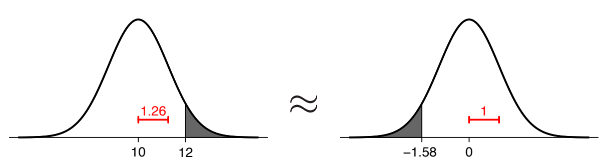


Example

Suppose X_1, X_2, ..., X_{10} are iid Normal(mean=10,SD=4)

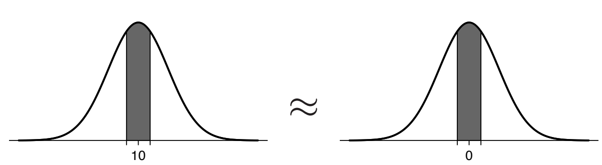
Then $\overline{X} \sim$ Normal(mean=10, SD $\approx$ 1.26). Let $Z = (\overline{X} - 10)/1.26$.

Pr($\overline{X} > 12$)?



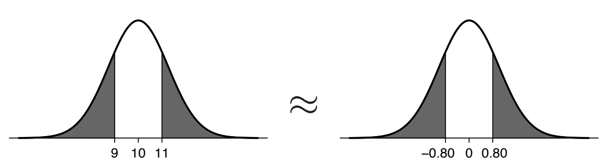
$\approx$ 5.7%

Pr($9.5 < \overline{X} < 10.5$)?



$\approx$ 31%

Pr($|\overline{X} - 10| > 1$)?



$\approx$ 43%

Central limit theorm

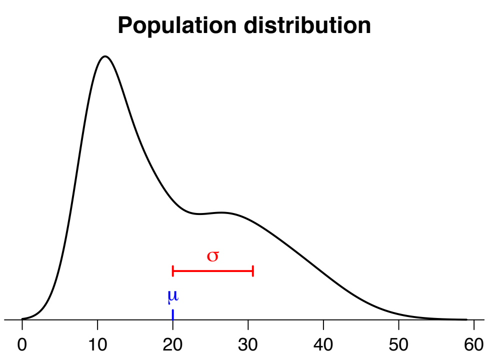
⟶ If X_1, X_2, ..., X_n are iid with mean μ and SD σ, and the sample size (n) is large, then

$$\overline{X} \text{ is approximately Normal}(\mu, \sigma/\sqrt{n}).$$

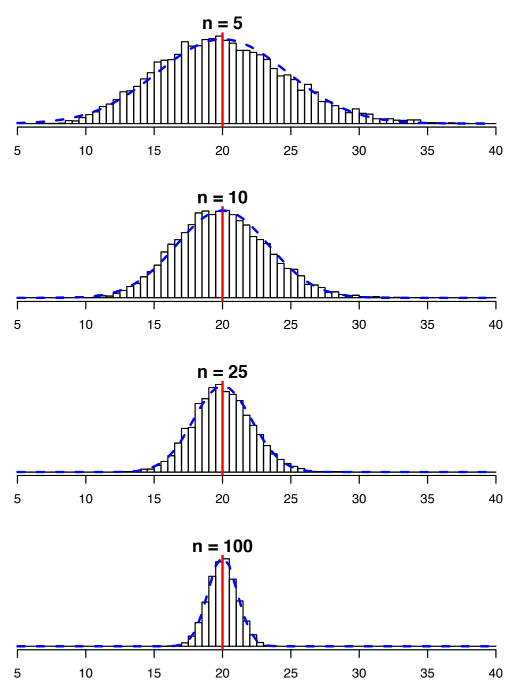
⟶ How large is large?

It depends on the population distribution.
(But, generally, not too large.)

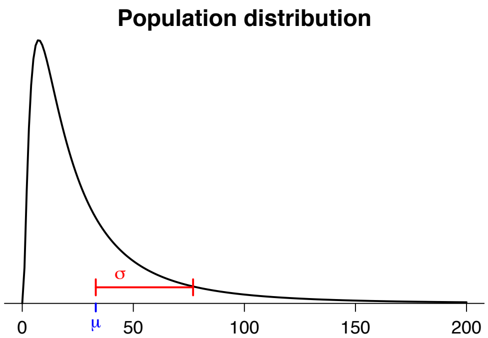
Example 1

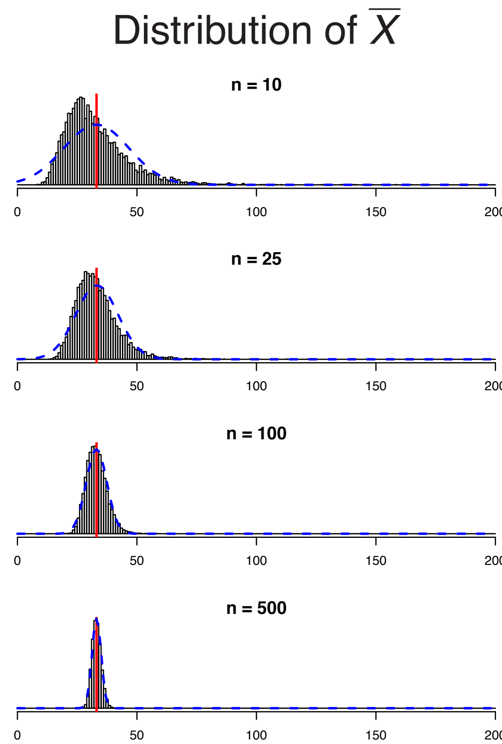


Distribution of $\overline{X}$

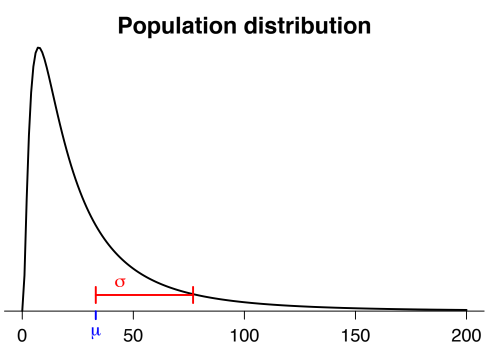


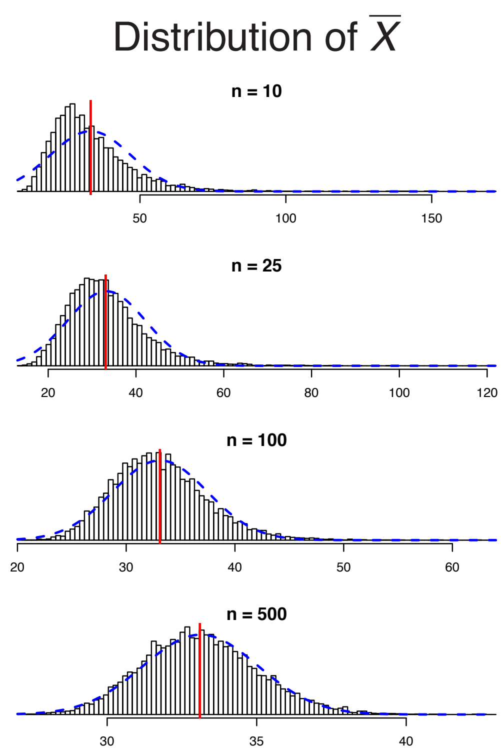
Example 2



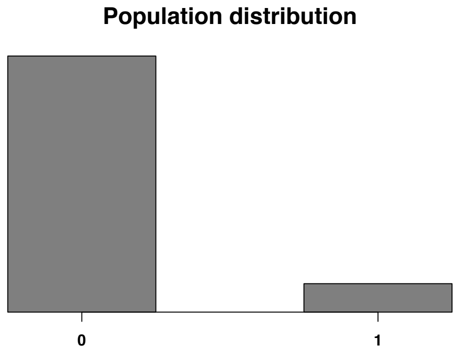


Example 2 (rescaled)





Example 3



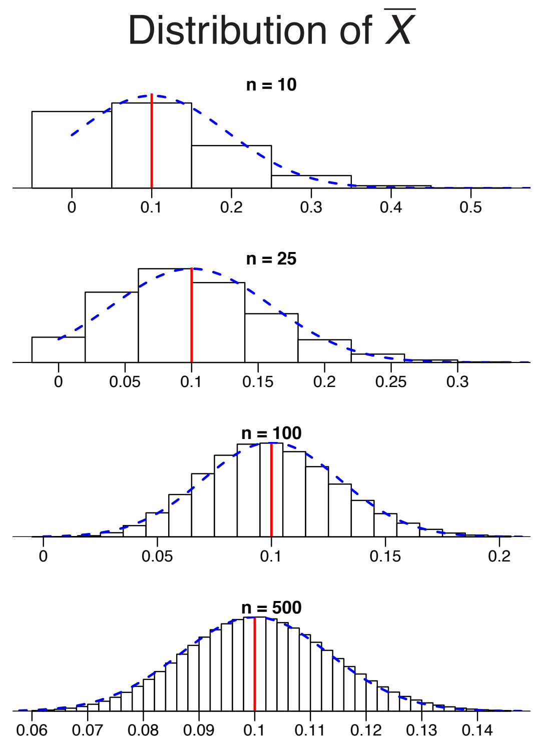
$\{X_i\}$ iid

$\Pr(X_i = 0) = 90\%$
$\Pr(X_i = 1) = 10\%$

$E(X_i) = 0.1$; $SD(X_i) = 0.3$

$\sum X_i \sim$ Binomial(n, p)

$\rightarrow$ $\overline{X}$ = proportion of 1's



The sample SD

$\longrightarrow$ Why use (n – 1) in the sample SD?

$$S = \sqrt{\frac{\sum (X_i - \overline{X})^2}{n - 1}}$$

$\longrightarrow$ If $\{X_i\}$ are iid with mean μ and SD σ, then

- $E(S^2) = \sigma^2$
- $E\{ \frac{n-1}{n} S^2 \} = \frac{n-1}{n} \sigma^2 < \sigma^2$

$\longrightarrow$ In other words:

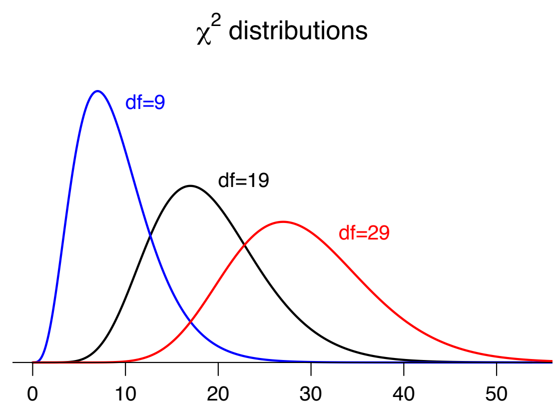
- $\text{Bias}(S^2) = 0$
- $\text{Bias}\left(\frac{n-1}{n} S^2 \right) = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2$

The distribution of the sample SD

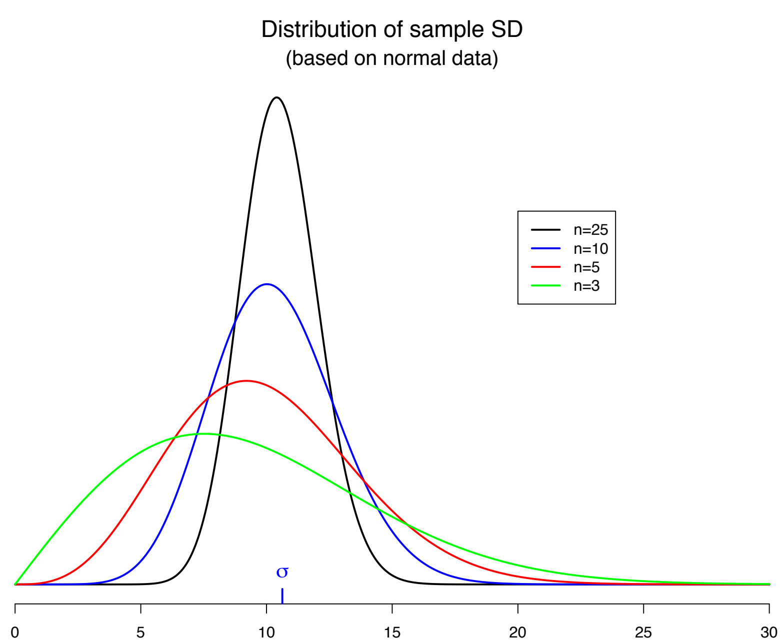
→ If $X_1, X_2, \ldots, X_n$ are iid Normal(μ, σ), then the sample SD S satisfies

$$(n-1)\ S^2/\sigma^2 \sim \chi^2_{n-1}$$

(When the X_i are not normally distributed, this is not true.)



Example



A non-normal example