

This work is licensed under a [Creative Commons Attribution-NonCommercial-ShareAlike License](https://creativecommons.org/licenses/by-nc-sa/4.0/). Your use of this material constitutes acceptance of that license and the conditions of use of materials on this site.

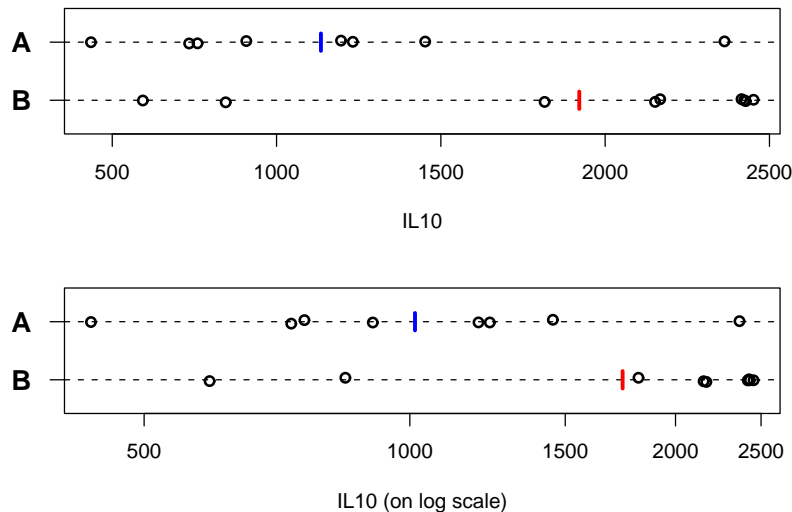


Copyright 2006, The Johns Hopkins University and Karl Broman. All rights reserved. Use of these materials permitted only in accordance with license rights granted. Materials provided "AS IS"; no representations or warranties provided. User assumes all responsibility for use, and all liability related thereto, and must independently review all materials for accuracy and efficacy. May contain materials owned by others. User is responsible for obtaining permissions for use from third parties as needed.

Example

Two strains of mice: A and B.

Measure cytokine IL10 (in males all the same age) after treatment.



Point: 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**.

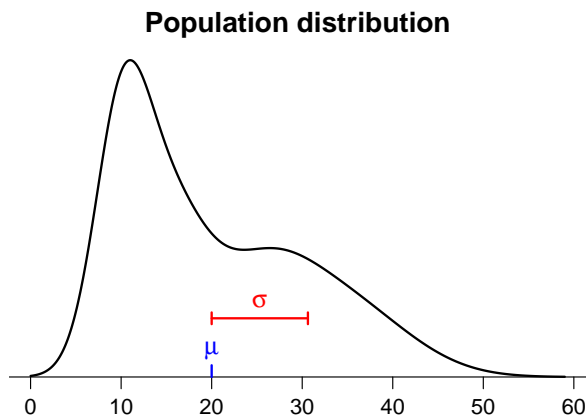
Examples:

- 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 cryptosporidium.
- 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

The object of our interest is 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, \dots, X_n independent and identically distributed (**iid**), following the population distribution.

Statistic: A numerical summary (function) of the X 's (that is, of the data). For example, the sample mean, sample SD, etc.

Estimator: A statistic, viewed as estimating some population parameter.
(estimate)

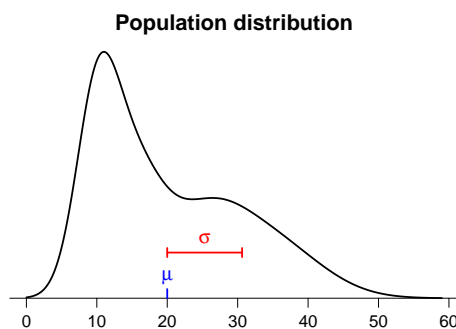
We write: $\hat{\theta}$ an estimator of θ $\bar{X} = \hat{\mu}$ an estimator of μ
 \hat{p} an estimator of p $S = \hat{\sigma}$ an estimator of σ

Parameters, estimators, estimates

- μ
 - The population mean
 - A **parameter**
 - A **fixed** quantity
 - Unknown, but what we want to know
- \bar{X}
 - The sample mean
 - An **estimator** of μ
 - A function of the data (the X 's)
 - A **random** quantity
- \bar{x}
 - The observed sample mean
 - An **estimate** of μ
 - A particular **realization** of the estimator, \bar{X}
 - A fixed quantity, but the result of a random process.

Estimators are random variables

Estimators have distributions, means, SDs, etc.

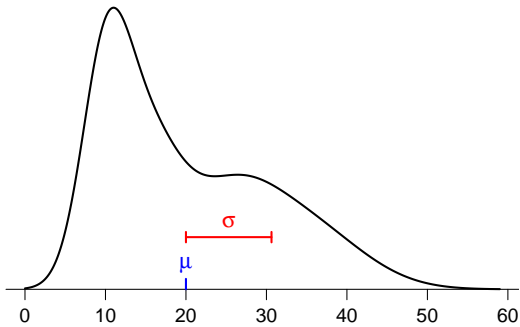


→ X_1, X_2, \dots, X_{10} → \bar{X}

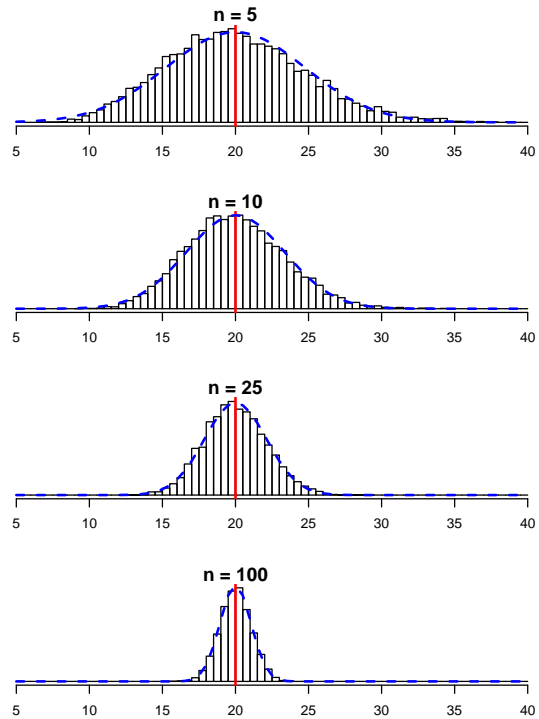
3.8	8.0	9.9	13.1	15.5	16.6	22.3	25.4	31.0	40.0	→	18.6
6.0	10.6	13.8	17.1	20.2	22.5	22.9	28.6	33.1	36.7	→	21.2
8.1	9.0	9.5	12.2	13.3	20.5	20.8	30.3	31.6	34.6	→	19.0
4.2	10.3	11.0	13.9	16.5	18.2	18.9	20.4	28.4	34.4	→	17.6
8.4	15.2	17.1	17.2	21.2	23.0	26.7	28.2	32.8	38.0	→	22.8

Sampling distribution

Population distribution



Distribution of \bar{X}

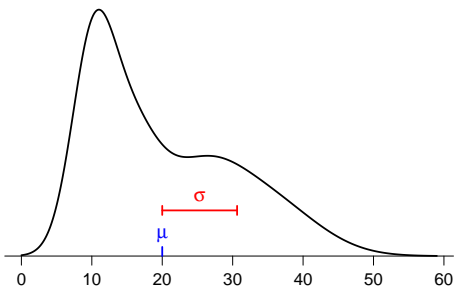


Sampling distribution depends on:

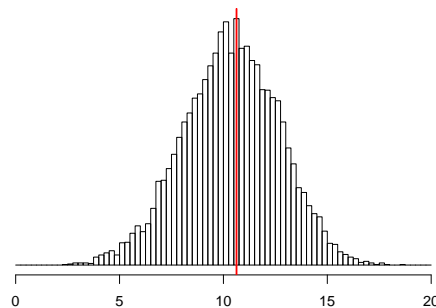
- The type of statistic
- The population distribution
- The sample size

Bias, SE, RMSE

Population distribution



Dist'n of sample SD (n=10)



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

Bias:

$$E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$$

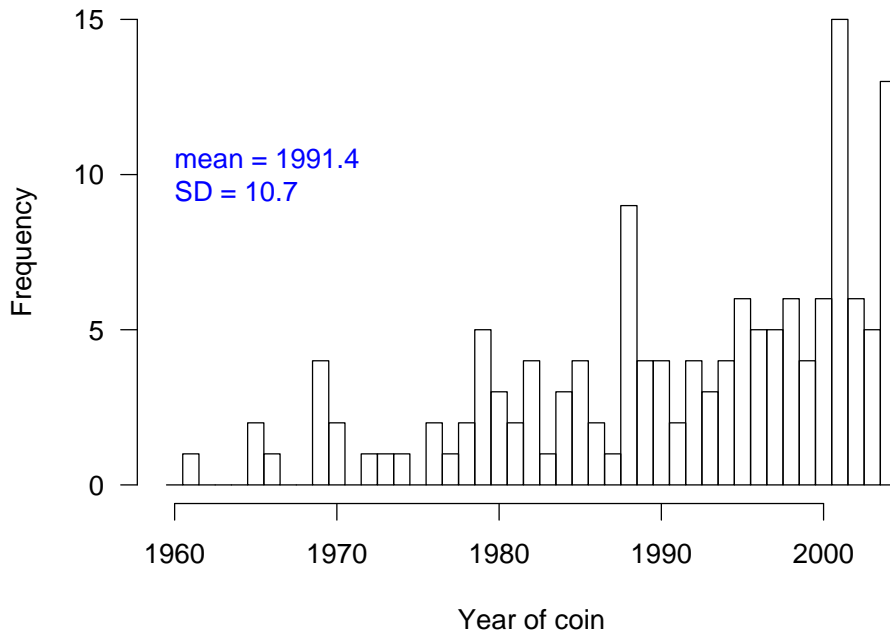
Standard error (SE):

$$SE(\hat{\theta}) = SD(\hat{\theta}).$$

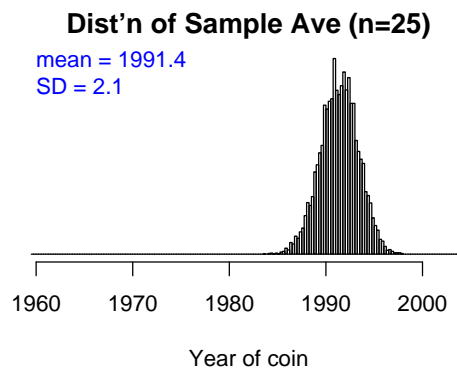
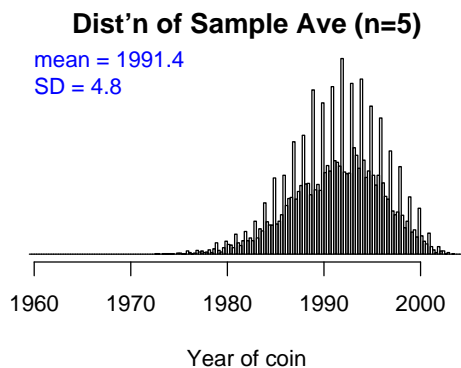
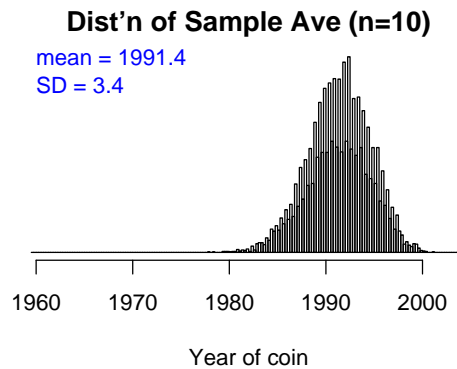
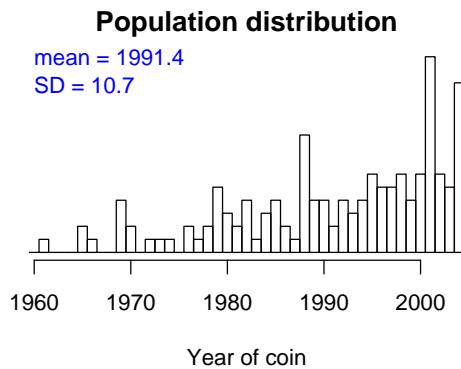
RMS error (RMSE):

$$\sqrt{E\{(\hat{\theta} - \theta)^2\}} = \sqrt{(\text{bias})^2 + (\text{SE})^2}.$$

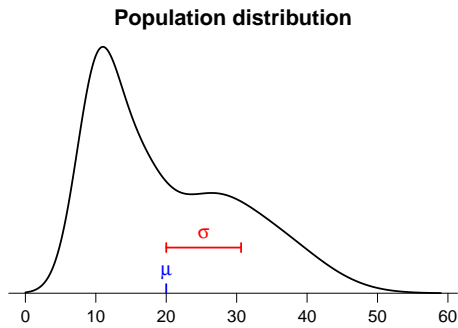
Example: years of coins



Example: years of coins



The sample mean



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

Mean of $\bar{X} = E(\bar{X}) = \mu$.

Bias = $E(\bar{X}) - \mu = 0$.

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

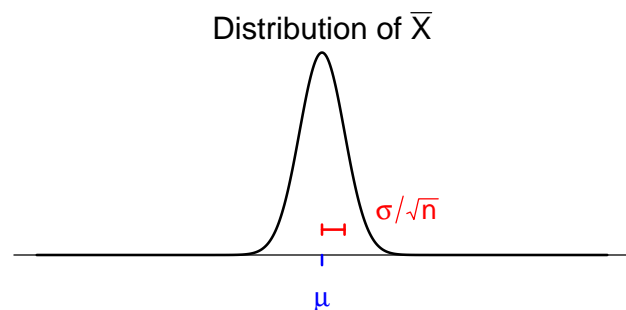
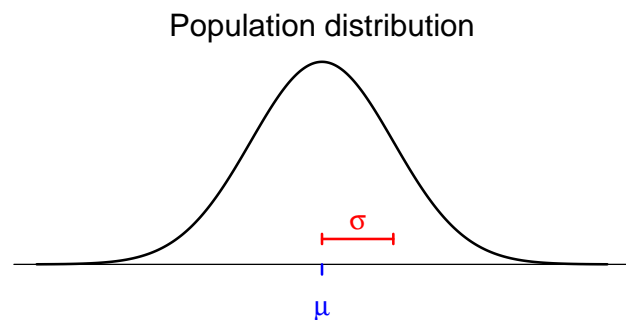
RMS error of $\bar{X} =$

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

If the population is normally distributed

If X_1, X_2, \dots, X_n are iid normal(μ, σ), then

$$\bar{X} \sim \text{normal}(\mu, \sigma/\sqrt{n}).$$

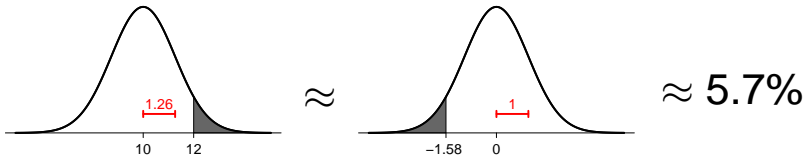


Example

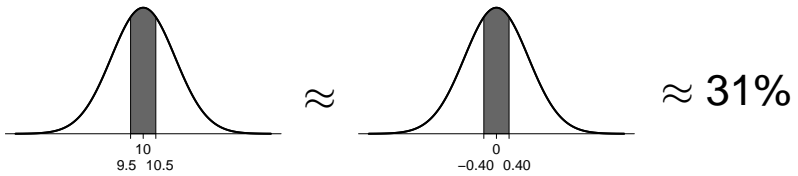
Suppose X_1, X_2, \dots, X_{10} are iid normal(mean=10,SD=4)

Then $\bar{X} \sim \text{normal}(\text{mean}=10, \text{SD} \approx 1.26)$; let $Z = (\bar{X} - 10)/1.26$.

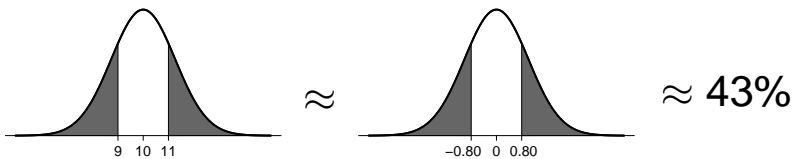
$\Pr(\bar{X} > 12)$?



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



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



Central limit theorem

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

and the sample size (n) is large,

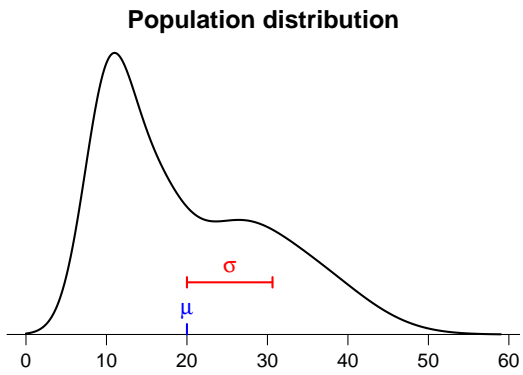
then \bar{X} is approximately normal($\mu, \sigma/\sqrt{n}$).

How large is large?

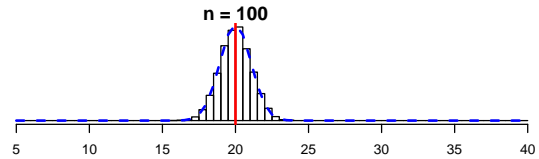
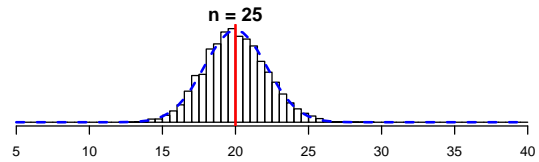
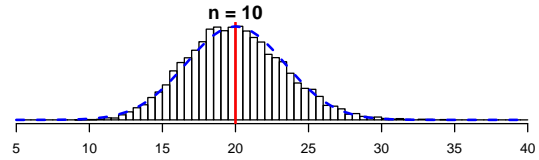
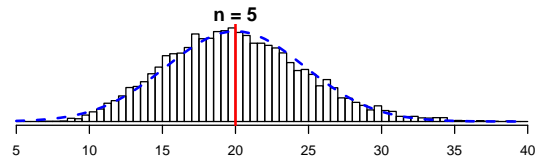
It depends on the population distribution.

(But, generally, not too large.)

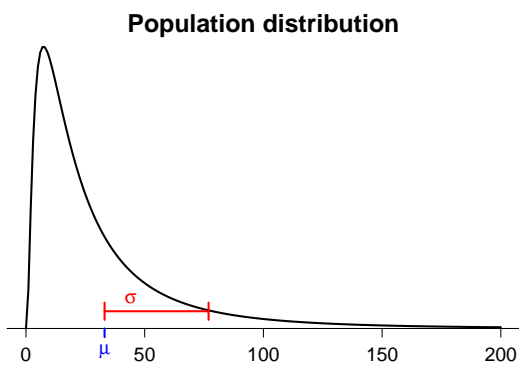
Example 1



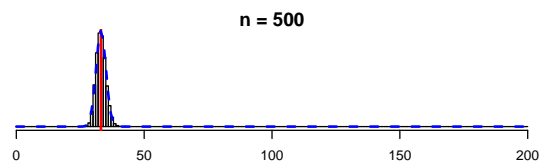
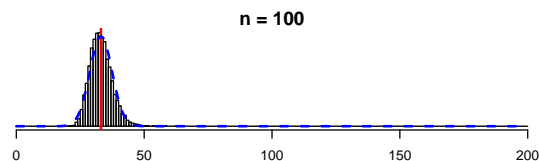
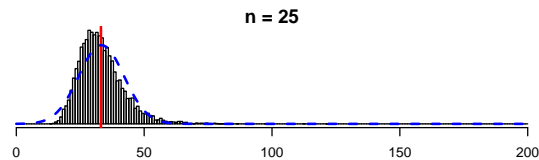
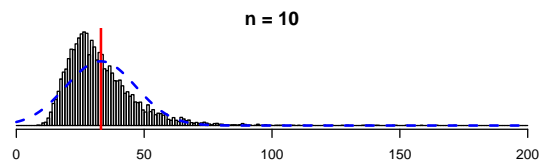
Distribution of \bar{X}



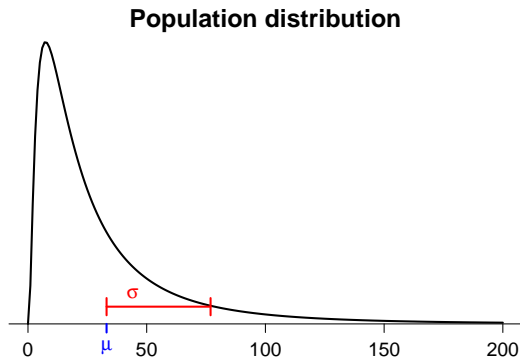
Example 2



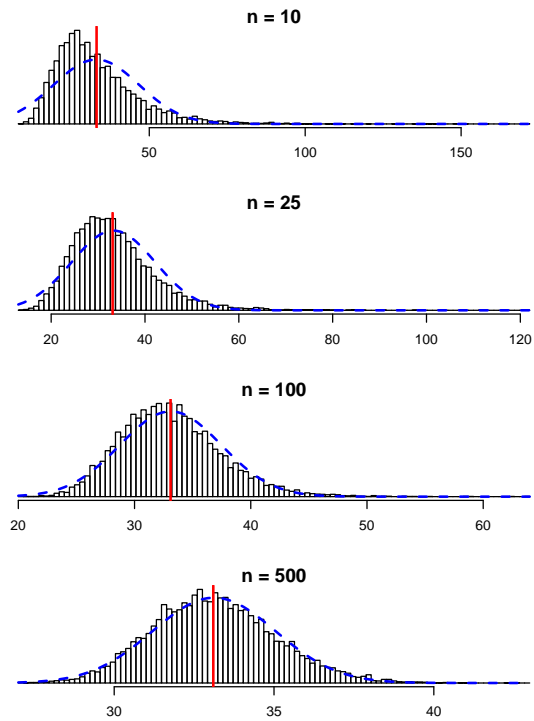
Distribution of \bar{X}



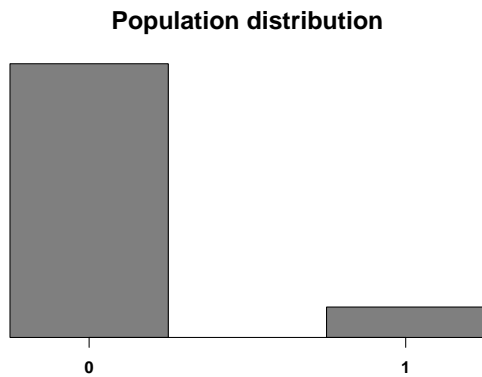
Example 2 (rescaled)



Distribution of \bar{X}



Example 3



$\{X_i\}$ iid

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

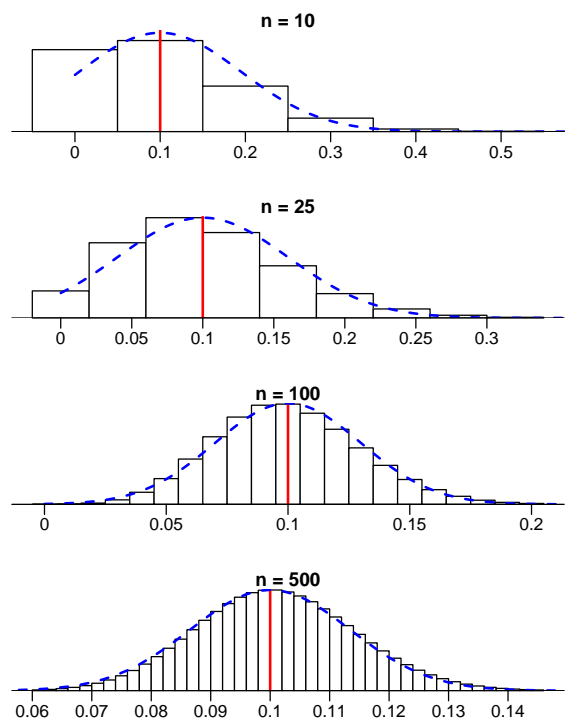
$$\Pr(X_i = 1) = 10\%$$

$$E(X_i) = 0.1; \text{SD}(X_i) = 0.3$$

$$\sum X_i \sim \text{binomial}(n, p)$$

\bar{X} = proportion of 1's

Distribution of \bar{X}



The sample SD

Why use $(n - 1)$ in the sample SD?

$$s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$$

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

$$E(s^2) = \sigma^2$$

$$\text{but } E\left\{ \frac{n-1}{n} s^2 \right\} = \frac{n-1}{n} \sigma^2 < \sigma^2$$

In other words:

$$\text{Bias}(s^2) = 0$$

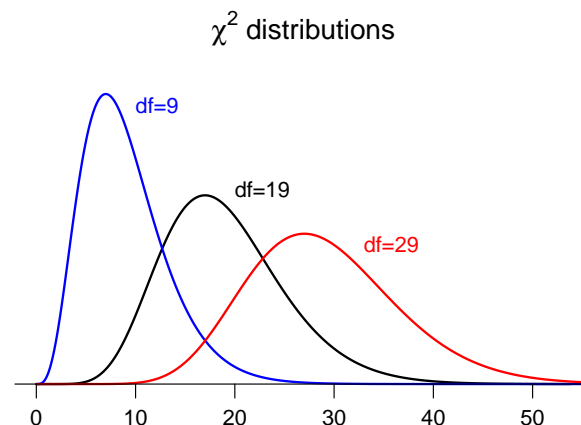
$$\text{but } \text{Bias}\left(\frac{n-1}{n} s^2\right) = \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{1}{n} \sigma^2$$

The distribution of the sample SD

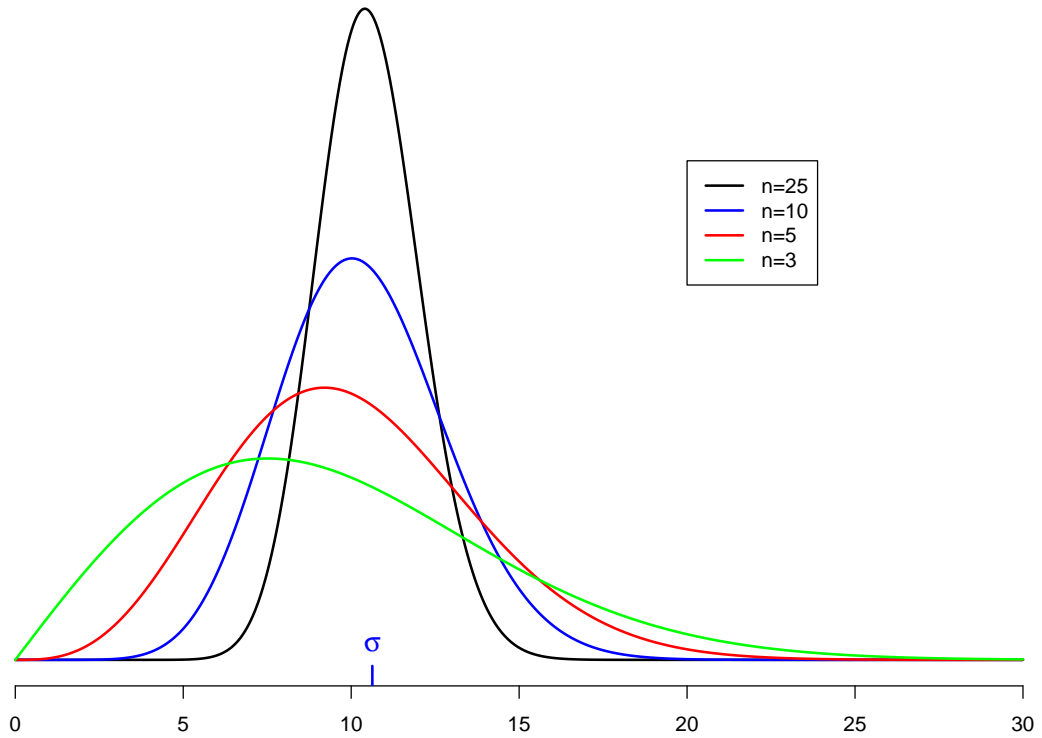
If X_1, X_2, \dots, X_n are iid normal(μ, σ)

then the sample SD, s , satisfies $(n - 1) s^2 / \sigma^2 \sim \chi_{n-1}^2$

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

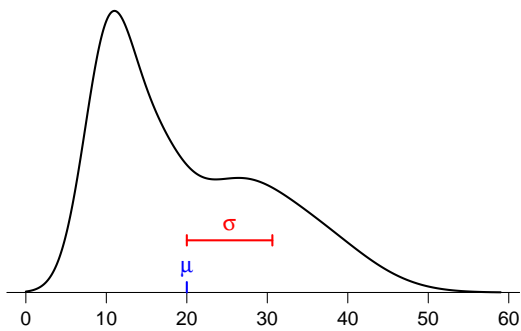


Distribution of sample SD (based on normal data)



A non-normal example

Population distribution



Distribution of sample SD

