Central Limit Theorem

Statistical Inference

The process of making guesses about the truth from a sample.

Sample statistics

$$\hat{\mu} = \overline{X}_n = \frac{\sum_{i=1}^{n} x}{n}$$

$$\hat{\sigma}^2 = s^2 = \frac{\sum_{i=1}^{n} (x_i - \overline{X}_n)^2}{n-1}$$

*hat notation ^ is often used to indicate "estimate"

Truth (not observable)

Population parameters

$$\mu = \frac{\sum_{i=1}^{N} x}{N} \qquad \sigma^2 = \frac{\sum_{i=1}^{N} (x_i - \mu)^2}{N}$$

Sample (observation)

Make guesses about the whole population

Statistics vs. Parameters

- **Sample Statistic** – any summary measure calculated from data (e.g., a mean, a difference in means or proportions, an odds ratio, or a correlation coefficient)
  - e.g., the mean vitamin D level in a sample of 100 men is 63 nmol/L

- **Population parameter** – the true value/true effect in the entire population of interest
  - e.g., the true mean vitamin D in all middle-aged and older European men is 62 nmol/L

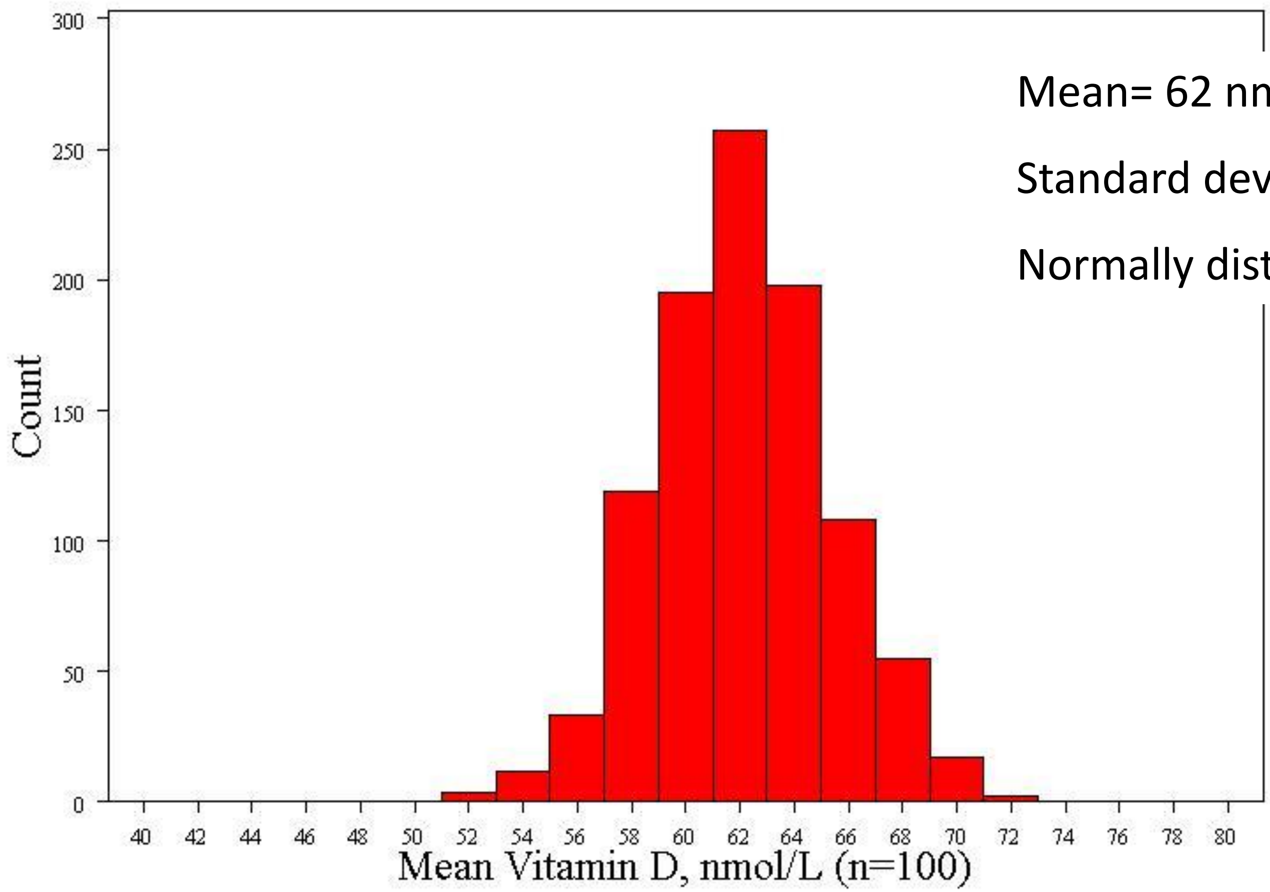
Distribution of a statistic

- Two approaches to determine the distribution of a statistic:
  - 1. Computer simulation
    - Repeat the experiment over and over again virtually!
    - More intuitive; can directly observe the behavior of statistics.
  - 2. Mathematical theory
    - Proofs and formulas!
    - More practical; use formulas to solve problems.

Distribution of the sample mean, computer simulation…

1. Specify the underlying distribution of vitamin D in all European men aged 40 to 79.
   - Right-skewed
   - Standard deviation = 33 nmol/L
   - True mean = 62 nmol/L (this is arbitrary; does not affect the distribution)
2. Select a random sample of 100 virtual men from the population.
3. Calculate the mean vitamin D for the sample.
4. Repeat steps (2) and (3) a large number of times (say 1000 times).
5. Explore the distribution of the 1000 means.

Distribution of mean vitamin D (a sample statistic)



Mean= 62 nmol/L

Standard deviation = 3.3 nmol/L

Normally distributed

Java applet

http://onlinestatbook.com/stat_sim/sampling_dist/

The Central Limit Theorem

If all possible random samples, each of size *n*, are taken from any population with a mean μ and a standard deviation σ, the sampling distribution of the sample means (averages) will:

1. have mean: $$\mu_{\bar{x}} = \mu$$

2. have standard deviation: $$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

3. be approximately normally distributed *regardless* of the shape of the parent population (normality improves with larger *n*).

$\mu_{\bar{x}}$ The mean of the sample means.

$\sigma_{\bar{x}}$ The standard deviation of the sample means.

Also called “standard error of the mean (SEM)”

Mathematical Proof

If *X* is a random variable from any distribution with known mean, E(*x*), and variance, Var(*x*), then the expected value and variance of the average of *n* observations of X is:

$$E(\overline{X}_n) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} E(x)}{n} = \frac{nE(x)}{n} = E(x)$$

$$Var(\overline{X}_n) = Var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right) = \frac{\sum_{i=1}^{n} Var(x)}{n^2} = \frac{nVar(x)}{n^2} = \frac{Var(x)}{n}$$