Lecture 2

Statistical background review (I)

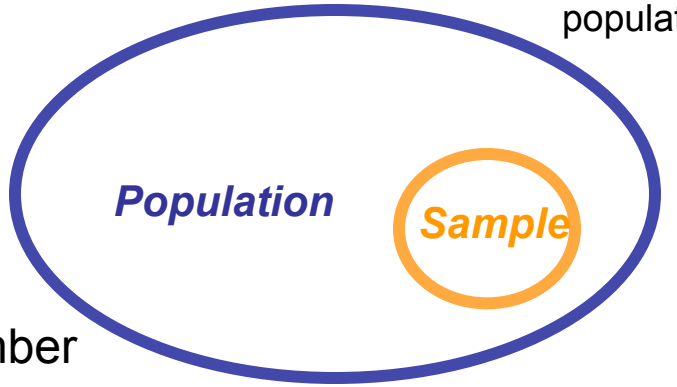
MCB 416A/516A
Statistical Bioinformatics and Genomic Analysis

Prof. Lingling An
Univ of Arizona

Population versus sample

- **Population:** The entire group of individuals in which we are interested but can't usually assess directly
  - Example: all male adults in Arizona

- **Sample:** A part of the population we actually examine and for which we do have data
  - Example: Take a sample of 1000 male adults in Arizona
    - How well the sample represents the population depends on the *sampling design*.



- A **parameter** is a number describing a characteristic of the population.
  - Example: average height of the male adults in Arizona
  - Usually parameter is unknown and need to be estimated by using ----

- A **statistic** is a number describing a characteristic of a sample.
  - ——Example: average height of these 1000 people
  - ——Different sample results in different statistic

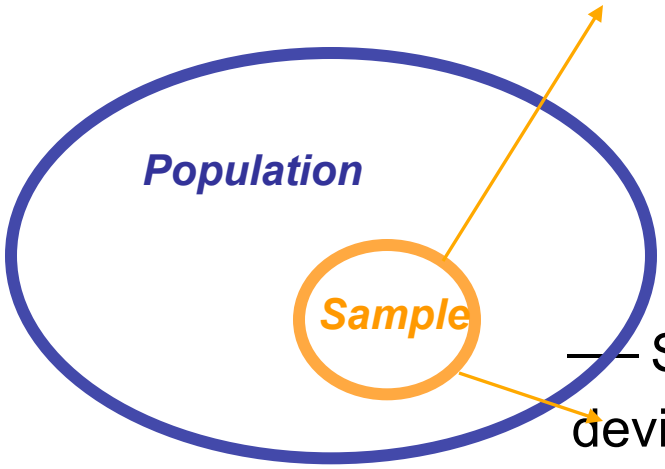
Parameter – statistic- estimate

More explanations:

- Parameter, which is unknown, is our interest,
- A statistic, distinct from an unknown parameter, can be computed from a sample.
- A statistic used to estimate a parameter is called an **estimate**.
  - For instance, the *sample mean* is a statistic and an estimate for the *population mean*, which is a parameter.
  - "Statistic" is more general than "estimate".
- Very often, we'll use the term "estimate" which is for a sample, corresponding to the "parameter" for a population.
- More specifically, "estimate" means point estimate.

Mean and standard deviation

- **Population:**
  - Population mean μ
  - Population standard deviation σ



- **Sample:**
  - — Sample mean

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i$$

  - — Sample standard deviation: variation around the mean

$$s = \sqrt{\frac{1}{n-1}\sum_{1}^{n}(x_i - \bar{x})^2} \quad \text{(unbiased)}$$

$$s_n = \sqrt{\frac{1}{n}\sum_{1}^{n}(x_i - \bar{x})^2} \quad \text{(biased)}$$

Normal Distribution

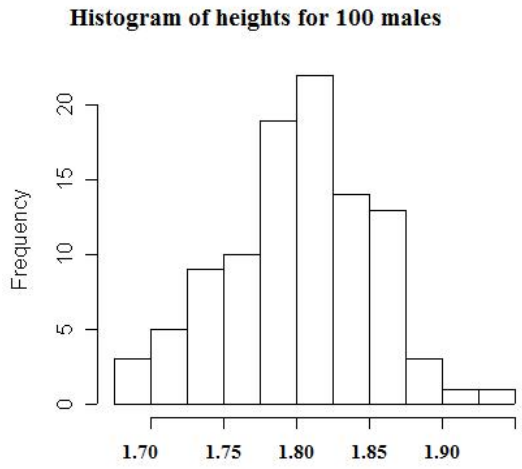
- Normal—or Gaussian—distribution is a symmetrical, bell- shaped density curve that describes the data clustering around its mean.

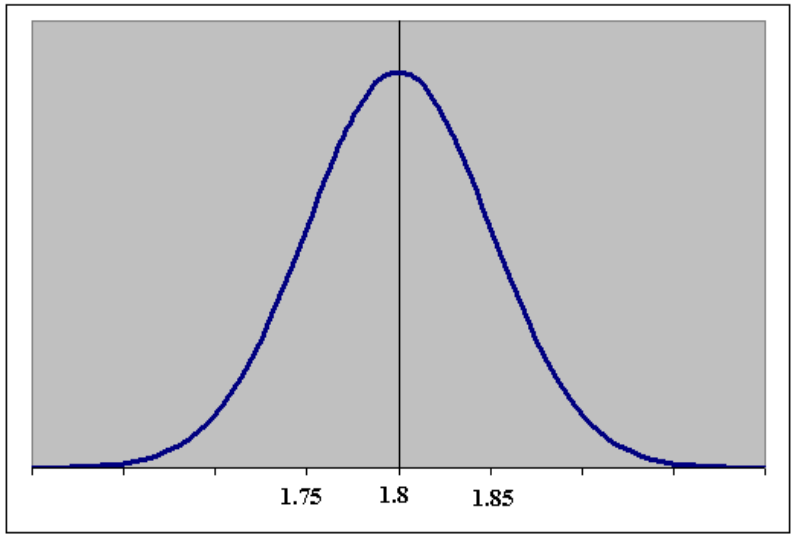
Example: heights of male adults in the United States are roughly normally distributed, with a mean of about 70 inches (1.8 m).

- Most men have a height close to the mean
- Small number of outliers have a height significantly above or below the mean.

Normal distribution (cont'd)

- use a sample, a histogram of male heights will appear similar to a bell curve, with the correspondence becoming closer if more data are used.

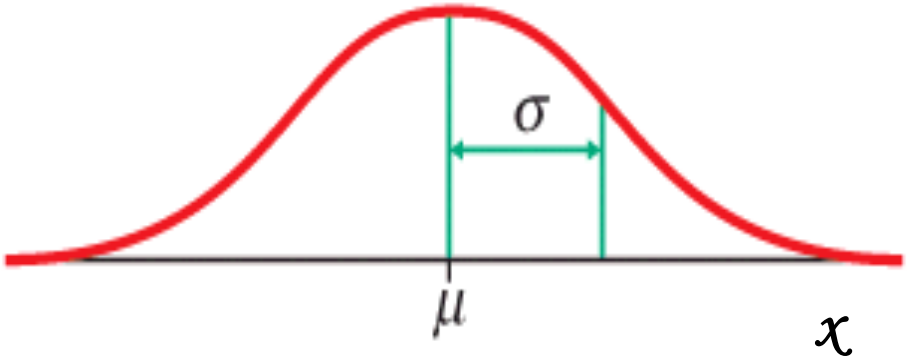


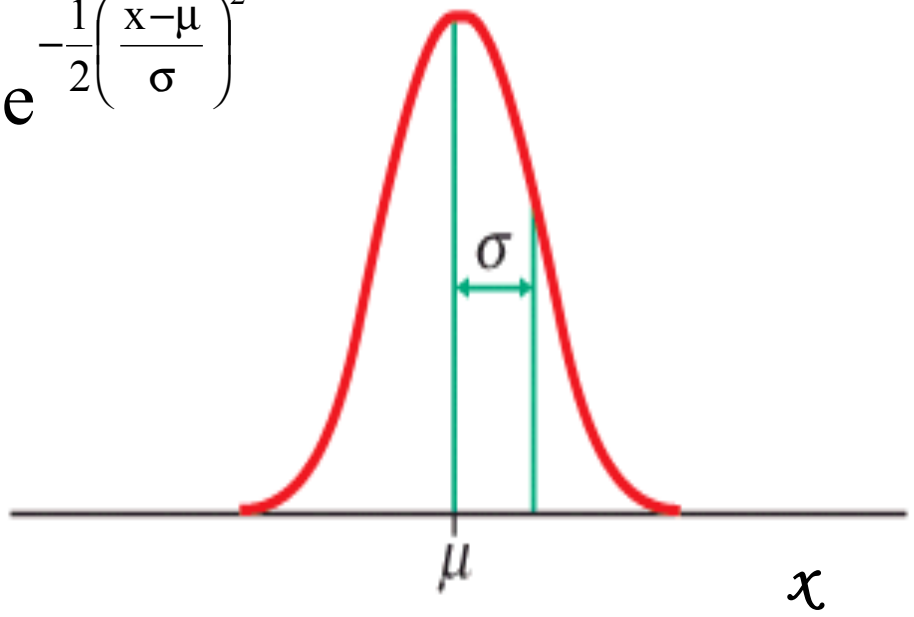


Normal distribution (cont'd)

- it's defined by a mean μ (*mu*, location) and a standard deviation σ (*sigma,* shape): $N(\mu, \sigma)$.

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





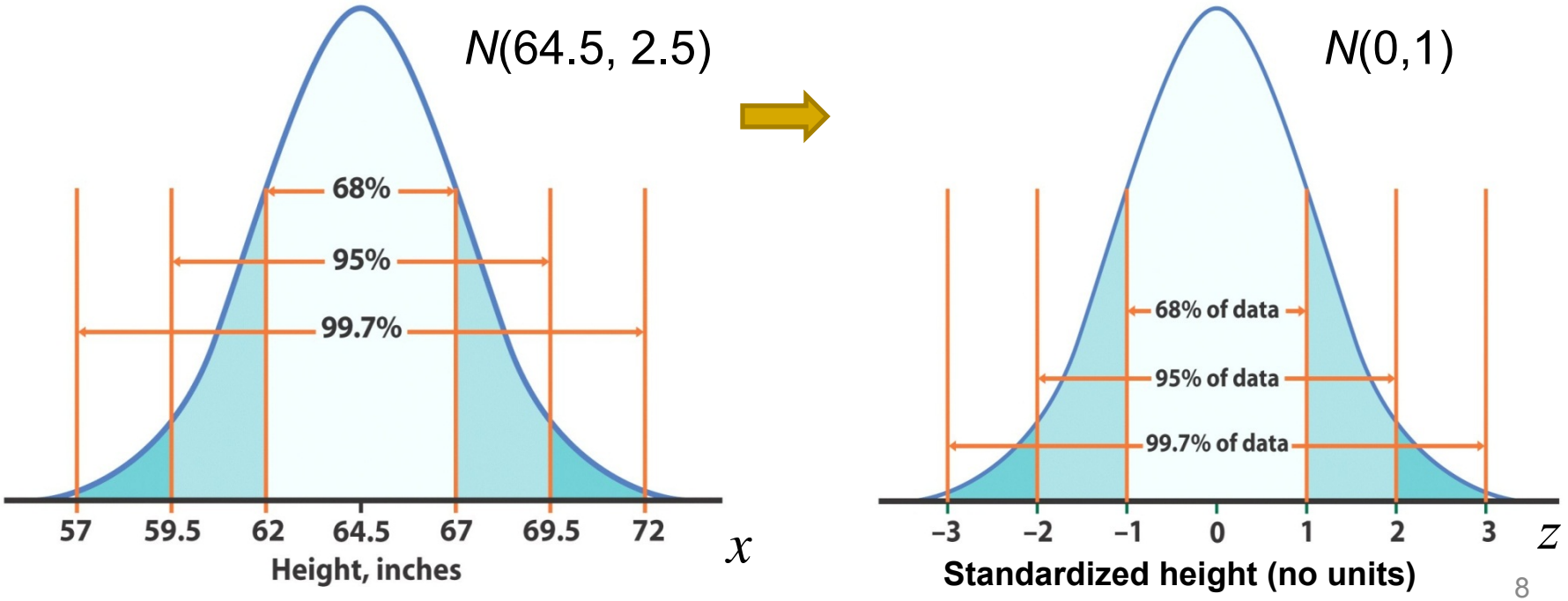
e = 2.71828… The base of the natural logarithm

π = pi = 3.14159…

The standard Normal distribution

Because all Normal distributions share the same properties, we can **standardize** our data to transform any Normal curve $N(\mu, \sigma)$ into the standard Normal curve $N(0,1)$ by ***shifting*** and ***scaling***.

For each x we calculate a new value, z (called a z-score).



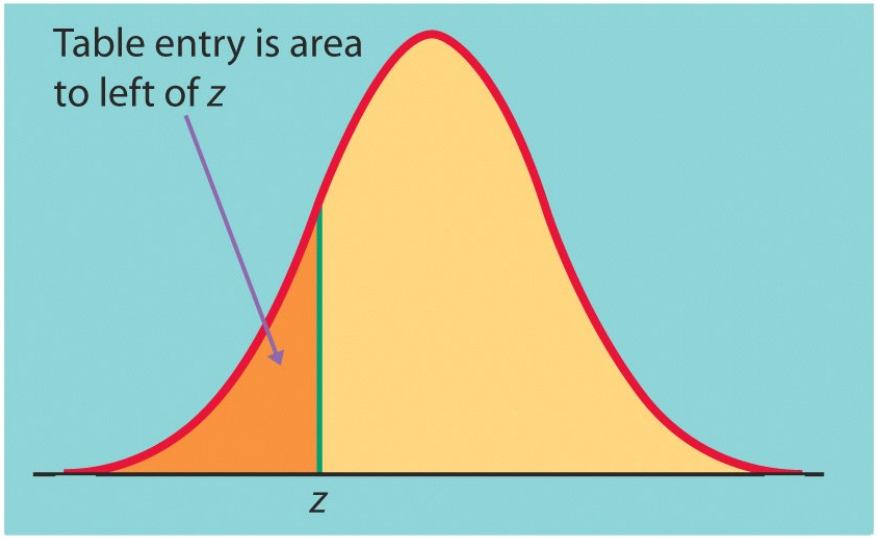
Standardizing: calculating z-scores

A ***z*-score** measures the number of standard deviations that a data value x is from the mean μ.

$$z = \frac{(x - \mu)}{\sigma}$$

THE STANDARD NORMAL TABLE

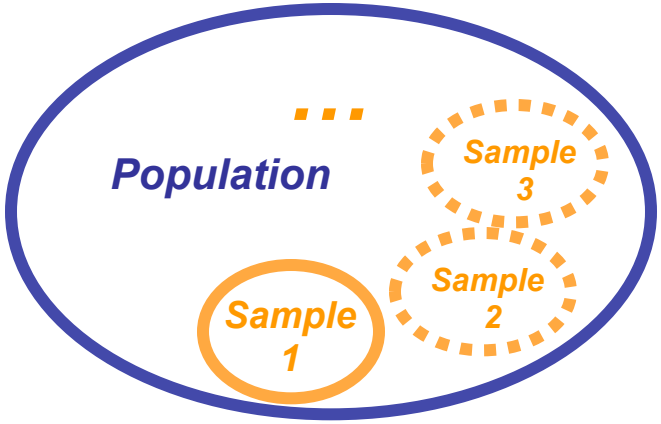
Table A is a table of areas under the standard Normal curve. The table entry for each value z is the area under the curve to the left of z.



Why confidence interval?

- We have discussed point estimates:
  - as an estimate of population mean μ
  - as an estimate of population standard deviation σ

- These point estimates are almost never exactly equal to the true values they are estimating.
- In order for the point estimate to be useful, it is necessary to describe just how far off from the true value it is likely to be.



Why confidence interval? (cont'd)

- Since the parameter (e.g., population mean) will not be exactly equal to the estimate (e.g., sample mean), , it is best to construct a **confidence interval around that is likely to cover the parameter**.

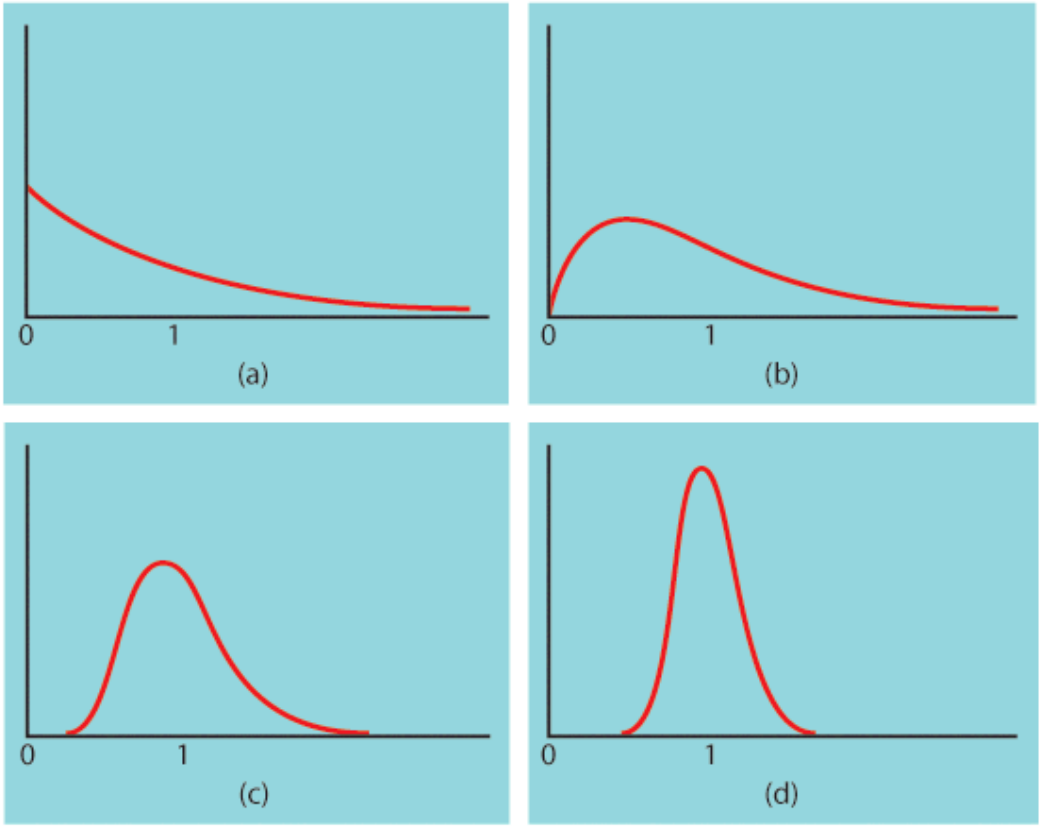
- We can then quantify our level of confidence that the parameter (population mean, as an example) is actually covered by the interval.

The central limit theorem

- Suppose we have a population described by a random variable *X* with a mean *μ and a* standard deviation σ.
  - Any population (normal, uniform, exponential, etc)
- Suppose we now take random samples from this population, each with a fixed and large sample size *n.*
  - Each sample will have a sample mean $\overline{x}$, and this $\overline{x}$ will not, in general, be equal to the population mean.
  - After repeated samplings, we will have built a population of $\overline{x}$*s . The* $\overline{x}$*s are* themselves random variable values and they have their own probability distribution.
- **Central Limit Theorem:** when *n* is large enough, the sampling distribution of $\overline{x}$ is approximately normal: $N(\mu, \sigma/\sqrt{n})$,.

The central limit theorem (illustration)

Population with strongly skewed distribution



Sampling distribution of $\bar{x}$ for n = 2 observations

Sampling distribution of $\bar{x}$ for n = 10 observations

Sampling distribution of $\bar{x}$ for n = 25 observations

The central limit theorem (cont'd)

- Note: for the population has normal distribution N(μ , σ), the sampling distribution of $\bar{x}$ is exactly normal: $N(\mu, \sigma/\sqrt{n})$, regardless what n is.

Confidence interval (general)

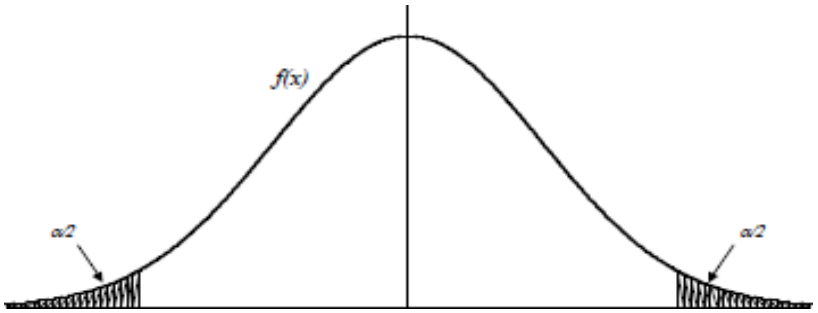
A level C confidence interval for a parameter has two parts:

- An interval calculated from the data, usually of the form

  Estimate ± critical value * Standard Error of the estimate

- A confidence level C, which gives the probability that the interval will capture the true parameter value in repeated samples, or the success rate for the method.

Confidence Interval on a Mean



- the (1–α)×100% confidence interval on the population mean:

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

where $z_{\alpha/2}$ is the *critical value* corresponding to a tail area of $\alpha/2$.

- If the population is not from a normal distribution, roughly we need n ≥ 30
- Online z-value: http://davidmlane.com/hyperstat/z_table.html

Exercises:

- A physical therapist studying muscular strength is willing to assume muscle strength scores are Normal with a standard deviation 12. A sample of 15 individuals demonstrates a mean muscular strength score of 84.3. Calculate a 95% confidence interval for μ

- (continued:) How large must *n* be for the width of the 99% confidence interval to be less than 5.0?

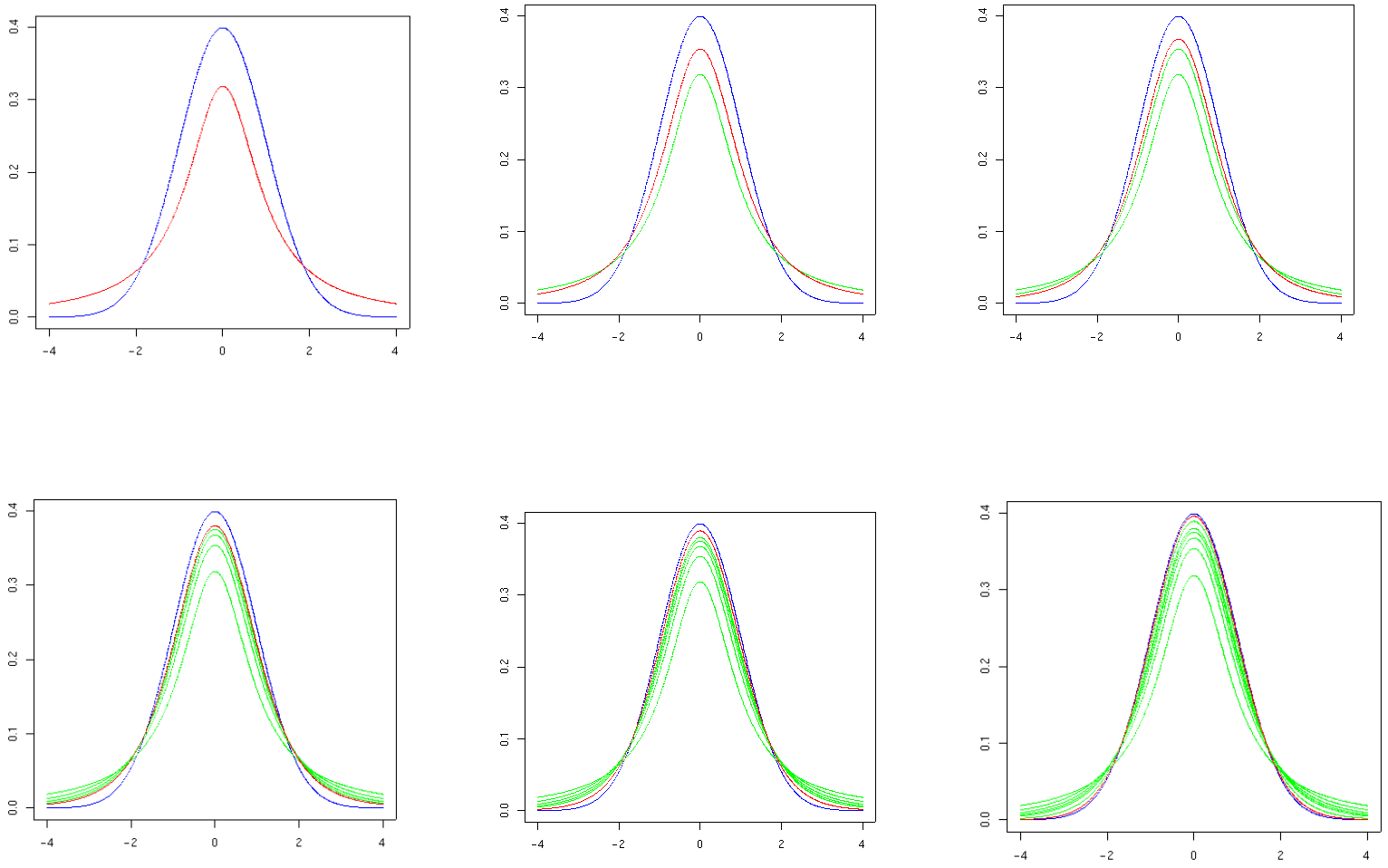
The *t* distribution – one sample t

- Why consider t-distribution?
  - —— we do not know the standard deviation of the population (very often!)
  - —— And sample sizes are small.
- then the sampling distribution of $\bar{x}$ follows a ***t* distribution with degrees of freedom *n* − 1 or equivalently,**

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} \quad , \mathrm{df} = n - 1$$

- so that we can use the standard t-table where df varies.

Density of the *t*-distribution (red and green) for 1, 2, 3, 5, 10, and 30 df compared to standard normal distr. (blue)



Confidence intervals when σ unknown

We have a set of data from a population with both μ and σ unknown. We use $\bar{x}$ to estimate μ, and s to estimate σ, using a t distribution (df= $n - 1$).

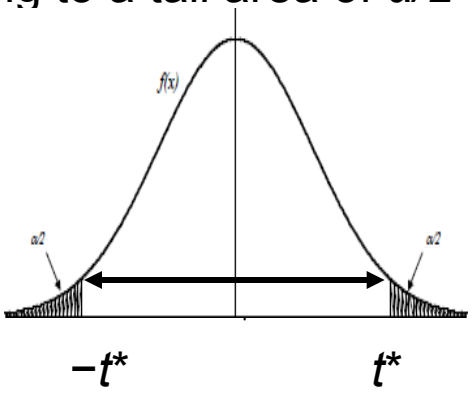
The (1–α)×100 % confidence interval for μ is:

$$\bar{x} \pm t^*_{(\alpha/2, n-1)} \frac{s}{\sqrt{n}}$$

where $t^*_{(\alpha/2,\ n-1)}$ is the *critical value* corresponding to a tail area of $\alpha/2$

Online t-table:

http://davidmlane.com/hyperstat/t_table.html



Exercise:

- ***Blood pressure.*** A study found a mean systolic blood pressure = 124.6 mm Hg in 35 individuals. The standard deviation *s* = 10.3 mm Hg.

  (A) Calculate the estimated standard error of the mean.

  (B) calculate the 90% confidence interval of the mean.

  (C) How many people would you need to study to decrease the standard error of the mean to 1 mm Hg? [Rearrange $se = s / \sqrt{n}$ to solve for n. Then plug-in values for *se* and *s*.]

Confidence Interval on Difference of Means (of two populations) (σ_1 and σ_2 KNOWN)

- Because we have two independent samples we use the difference between both sample averages ($\bar{x}_1 - \bar{x}_2$) to estimate ($\mu_1 - \mu_2$).
- The (1–α)×100% confidence interval on a difference in means:

$$(\bar{x}_1 - \bar{x}_2) \pm z_{\alpha/2}\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

  where n_1 and n_2 are the sample size from two samples, respectively, and $z_{\alpha/2}$ is the *critical value* corresponding to a tail area of *α/2.*

- This relationship is exact if the two populations are normally distributed. Otherwise, the confidence interval is approximately valid for large sample sizes ($n_1 \geq 30$ *and* $n_2 \geq 30$).

Confidence Interval on Difference of Means (σ_1 and σ_2 UNKNOWN but equal)

- If random samples of size n_1 *and* n_2 are drawn from two normal populations with equal but unknown variances, a 100(1-α) % confidence interval on the difference between the population means, $\mu_1 - \mu_2$ is:

$$(\bar{x}_1 - \bar{x}_2) \pm t_{(\alpha/2, n_1+n_2-2)} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

where S_p is a "pooled" estimator of the unknown standard deviation and is calculated as:

$$S_p = \sqrt{\frac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1 + n_2 - 2}}$$

- But this only be used if both populations are ***normally distributed***

Confidence Intervals on Paired Samples

- For the (1–α)×100% confidence interval on d *(difference between* a paired samples):

$$\bar{d} \pm t_{(\alpha/2,n-1)} * \frac{s_d}{\sqrt{n}}$$

- But this can only be used if both populations are ***normally distributed.***

Example:

- Suppose that ten identical twins were reared apart and the mean difference between the high school GPA of the twin brought up in wealth and the twin brought up in poverty was 0.07. If the standard deviation of the differences was 0.5, find a 95% confidence interval for the difference. Assume the distribution of GPA's is approximately normal.