We now move into inferential statistics, the "science" of making inferences and drawing conclusions about a population from a random sample of that population. Those inferences will be based on the Central Limit Theorem that tells us important things about the sampling distribution of means (and later for things related to means).

1. Given a random variable $y$ with an unknown type of distribution but with a mean $\mu = 8$ and a standard deviation $\sigma = 2$, what is $P(\bar{y} < 8.4)$ for the $\bar{y}$ in a random sample of 100 values of the variable $y$?

A 95% Confidence Interval for the Mean, $\mu$, of a Random Variable for a Sample of Size $n \geq 30$

Given a random variable $y$ with a mean $\mu$ and a standard deviation $\sigma$, then for a sample of size $n \geq 30$, $\bar{y}$ has a normal distribution with mean $\mu$ and standard deviation $\frac{\sigma}{\sqrt{n}}$ and $z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}}$ has a standard normal distribution with mean 0 and standard deviation 1. Therefore $P(-1.96 < Z < 1.96) = .95$ so there is a probability of .95 (95%, 1 chance in 20 of being wrong) that:

$$-1.96 < z < 1.96$$

and therefore that

$$-1.96 < \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}} < 1.96$$

since $z = \frac{\bar{y} - \mu}{\frac{\sigma}{\sqrt{n}}}$, which by simple algebra gives us the same probability (.95) that

$$-1.96 \frac{\sigma}{\sqrt{n}} < \bar{y} - \mu < 1.96 \frac{\sigma}{\sqrt{n}}$$

$$-\bar{y} - 1.96 \frac{\sigma}{\sqrt{n}} < -\mu < -\bar{y} + 1.96 \frac{\sigma}{\sqrt{n}}$$

$$\bar{y} + 1.96 \frac{\sigma}{\sqrt{n}} > \mu > \bar{y} - 1.96 \frac{\sigma}{\sqrt{n}}$$

and thus

$$\bar{y} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 1.96 \frac{\sigma}{\sqrt{n}}$$

This last interval, $\bar{y} - 1.96 \frac{\sigma}{\sqrt{n}} < \mu < \bar{y} + 1.96 \frac{\sigma}{\sqrt{n}}$, is called a 95% confidence interval for $\mu$. This means that the probability that such an interval includes the true value of $\mu$ for the population is .95. For $n \geq 30$ the sample variance, $s^2$, is a good approximation of $\sigma^2$ (that is, $s = \sigma$). So the interval

$$\bar{y} - 1.96 \frac{s}{\sqrt{n}} < \mu < \bar{y} + 1.96 \frac{s}{\sqrt{n}}$$

is also called a 95% confidence interval for the true value of $\mu$. 
2. Find a 95% confidence interval for $\mu$ from a sample of size 400 for the random variable $y$, assuming $\bar{y} = 8$ and $s = 2$.

3. The value $\bar{y} - 1.96 \frac{\sigma}{\sqrt{n}}$ is called the lower confidence limit for $\mu$. The value $\bar{y} + 1.96 \frac{\sigma}{\sqrt{n}}$ is called the upper confidence limit for $\mu$. If a different random sample of size 400 was drawn for $y$ in #1 and its mean $\bar{y}$ and its standard deviation, $s$, calculated, which would change: the true value of $\mu$ for the population or the values of the upper and lower confidence limits for the 95% confidence interval?

4. In light of the answer to #3 is it better to say: "there is a 95% chance that $\mu$ lies between the upper and lower confidence limits" or "the upper and lower confidence limits form an interval that has a probability of 95% of including the true value of $\mu"$?

---

**Geometrical Interpretation of a 95% Confidence Interval**

Consider the following drawing of a the graph of the pdf for a standard normal distribution $Z$ with mean 0 and standard deviation 1:

The shaded areas represent the regions for $Z < -1.96$ and for $Z > 1.96$. Their area is $P(Z < -1.96) = .025$ and $P(Z > 1.96) = .025$. The total area of the two shaded regions is .05 ($= .025+.025$). The area under the pdf curve that is not shaded is .95 ($= 1 - .05$). We label these points $z_{.025}$ and $-z_{.025}$. That is, $z_{.025} = 1.96$. The confidence interval is thus $z_{.025} < Z < z_{.025}$ or in terms of $\mu$, $\bar{y} - z_{.025} \frac{s}{\sqrt{n}} < \mu < \bar{y} + z_{.025} \frac{s}{\sqrt{n}}$. We know that

$$P\left(\bar{y} - z_{.025} \frac{s}{\sqrt{n}} < \mu < \bar{y} + z_{.025} \frac{s}{\sqrt{n}}\right) = .95 = 1 - 2(.025)$$