Statistical inference: confidence intervals

- **confidence**
  property associated with the (typically high) probability that an estimate in the form of a range of values that claims to contain the parameter being estimated

- **confidence interval**
  an estimate of a population parameter in the form of a best **estimate** with a **margin of error** (*ME*), producing a range of values; a confidence interval has the form

$$estimate \pm ME$$

- **confidence level**
  the rate of successful estimation in the sampling method (in practice, we wish this to be as high as possible)

Confidence intervals for $\hat{p}$

Recall that the sampling distribution for the sample proportion $\hat{p}$ in a SRS selected from the population and chosen to satisfy the 10% Condition and the Success/Failure Condition follows the $N\left(p, \sqrt{\frac{pq}{n}}\right)$ distribution.

Because of the Central Limit Theorem, roughly 95% of all samples will result in values of $\hat{p}$ that lie within 2 standard deviations of the true mean value p. This means that 95% of the time, p lies within 2 standard deviations of any sampled value of $\hat{p}$.

Since two standard deviations in the sampling distribution of $\hat{p}$ equals $2SD(\hat{p}) = 2\sqrt{\frac{pq}{n}}$, a quantity which is unknown to us (since p is unknown), we must estimate it as $2SE(\hat{p}) = 2\sqrt{\frac{\hat{p}\hat{q}}{n}}$. This last quantity our a **margin of error** for the **estimate** $\hat{p}$ to the (unknown) value of p.

Therefore,

$$\hat{p} \pm 2SE(\hat{p})$$

is a **95% confidence interval** for p: In 95% of all SRS the two numbers $\hat{p} - 2SE(\hat{p})$ and $\hat{p} + 2SE(\hat{p})$ will bound an interval that contains the true value of p.

More generally, when we choose a **confidence level** C (like $C = .95$), an associated **critical value** z^* of the standardized normal variable z can be found so that the central probability of finding z between $-z^*$ and z^* equals C: $P(-z^* \le z \le z^*) = C$. (For $C = .95$, we have $z^* = 1.96 \approx 2$.)

Therefore, a proportion C of all samples produce values of $\hat{p}$ that lie within z^* standard deviations of p, or rather, p lies within z^* standard deviations of any one value of $\hat{p}$ with a confidence level measured by C.

The margin of error for the estimate $\hat{p}$ of the (unknown) value of p is $ME = z^* \cdot SE(\hat{p}) = z^* \sqrt{\dfrac{\hat{p}\hat{q}}{n}}$.

So

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

is a **level** C **confidence interval for** μ. In a proportion C of all possible samples, the two numbers $\hat{p} - z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$ and $\hat{p} + z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$ bound an interval that contains the true value of μ.

[TI-83: `STAT TESTS 1-PropZTest…`]

Choosing the sample size

The margin of error in using $\hat{p}$ as an estimate for p is

$$ME = z^* \sqrt{\frac{\hat{p}\hat{q}}{n}}$$

To solve for the sample size n, we would need to know the value of $\hat{p}$ before we even know the size of the same we will select to measure $\hat{p}$! We can, however, estimate it. If we use the estimate p^*, then solving for n yields

$$n = \left(\frac{z^*}{m}\right)^2 p^*(1 - p^*)$$

Using the conservative estimate $p^* = 0.5$ produces a larger value of n in most circumstances (but larger samples will produce even more accurate estimates, so this need not be bad).