7.2 Finding a confidence interval for the sample proportion

- in practice we do not know p, we use the point estimate $\hat{p}$ which is calculated from the sample.
- for large samples we know that p is $N\left(p, \sqrt{\frac{p(1-p)}{n}}\right)$ by the Central Limit Theorem.
- we can find the z-score for the desired confidence level. we denote the confidence level $1-\alpha$, which gives the error level α. we use the equation

$$2\Phi(z) - 1 = 1 - \alpha$$

  to find the z-score.
- confidence intervals are given as

$$\text{point estimate} \pm \text{margin of error}$$

- the exact margin of error is given as

$$z\sqrt{\frac{p(1-p}{n}}$$

- the confidence interval for the sample proportion is given as

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$$

- ex the following question was asked on the GSS "would you be willing to pay higher prices to help the environment?" $n = 1154$, yes $= 518$.
  (a) construct a 95% confidence interval for the population proportion for those responding 'yes'.
  (b) interpret this interval.
- ex construct a confidence interval for those answering 'no'.
- question : what sample size is needed for the normality assumption?
  in general you should have at least 15 successes and 15 failures so that

$$n\hat{p} \geq 15 \qquad \text{and} \qquad n(1-\hat{p}) \geq 15$$

- we can also construct other intervals for different levels of confidence.
- ex we have data on the following question : "is it ok for a husband to refuse to have children if the wife wants to have children?" $n = 568$, yes $= 366$, no $= 232$.
  (a) find a 99% confidence interval for the 'yes's

(b) find a 95% confidence interval for the 'yes's

(c) compare the two

- what is the error associated w/the confidence interval?

- summary : for the sample proportion we have the point estimate $\hat{p}$ that is calculated from the data. then the confidence interval is given as

$$\hat{p} \pm z\sqrt{\frac{\hat{p}(1-\hat{p}}{n}}$$

so long as we meet the following criteria

(a) data is obtained through randomization

(b) sample is large enough to use CLT

- interpretation : the confidence interval refers to the long run behavior of taking many like samples. in the long run, if many samples are taken we expect that the 95% confidence intervals will contain the true parameter 95% of the time.

7.3 Confidence intervals for the mean

- use the same principle as before

$$\text{point estimate} \pm \text{margin of error}$$

- for quantatitive data we have $\hat{\mu} = \bar{x}$ and $\hat{\sigma} = s$, where $se = s/\sqrt{n}$ for the sample mean.

- ex : from the GSS we have the question "how much tv do you watch?" software reported the following about the data

var	N	$\bar{x}$	s	se	95% CI
TV	905	2.983	2.361	0.0785	(2.83,3.14)

(a) what do the mean and st dev tell us about the distribution of the sample?

(b) how did the software get the st err? what does it mean?

(c) interpret the confidence interval given here.

- how is the margin of error found for small sample sizes?

  we use what is called the t distribution. in practice we don't know the population st dev, so we estimate it using s. (not needed for the proportion) when s is used, we need to use values from the t distribution. combination of the unknown variance and small sample sizes.

- t values are typically larger than z values.

- t values approach z values as n gets larger.
- conf intervals are larger b/c of this
- using the t distirubtion forces us to assume that the underlying distribution is approximately normally distributed.
- properties of the t distribution
  - bell shaped and symmetric about 0
  - properties depend on the degrees of freedom. the t distribution has a different shape for each of the degrees of freedom.
  - thicker tails and more spread out than the standard normal. this means that extreme values are more likely.
  - t score times se gives the margin of error for a confidence interval about the mean.
- using t to construct a confidence interval. the interval is given as

  $$\bar{x} \pm t\left(\frac{s}{\sqrt{n}}\right)$$

  so long as

  (a) data obtained through randomization

  (b) observations are approximately normal distribution
- ex eBay auction for Palm Handheld computers. we are given the following figures for sales of Palm's $x = (235, 225, 225, 240, 250, 250, 210)$.

  (a) check the assumptions.

  (b) find a 90% and a 95% confidence interval for the mean sales price.
- using the t score is robust for the normality assumption. this means that even if the normality assumption is wrong the t score is still good to use assuming the data hasn't been corrupted by large outliers.
- remember that for large values of n the t distribution is the same as the standard normal.
- always use t when σ is unknown and estimated. if σ is known, tables for normal distribution may be used if you believe the observations are normally distributed.