Sessions 3-4

Estimation of population parameters

Class Outline

- How to estimate the population parameters such as mean and proportion using a sample?
- How to adjust the estimate if the population standard deviation is not known?
- What should be the sample size for a desired bound on the margin of error?

Estimation of population mean μ: Example

Example: Credit card launch

- A university with 100,000 alumni is thinking of offering a new affinity credit card to its alumni.
- Profitability of the card depends on the average balance maintained by the cardholders.
- A market research campaign is launched, in which about 140 alumni accept the card in a pilot launch.
- Average balance maintained by these is \$1990 and the standard deviation is \$2833. Assume that the population standard deviation is \$2500 that was derived from previous launches.
- What can we say about the average balance that will be held after a full-fledged market launch?

ISB

Interval Estimates of Parameters

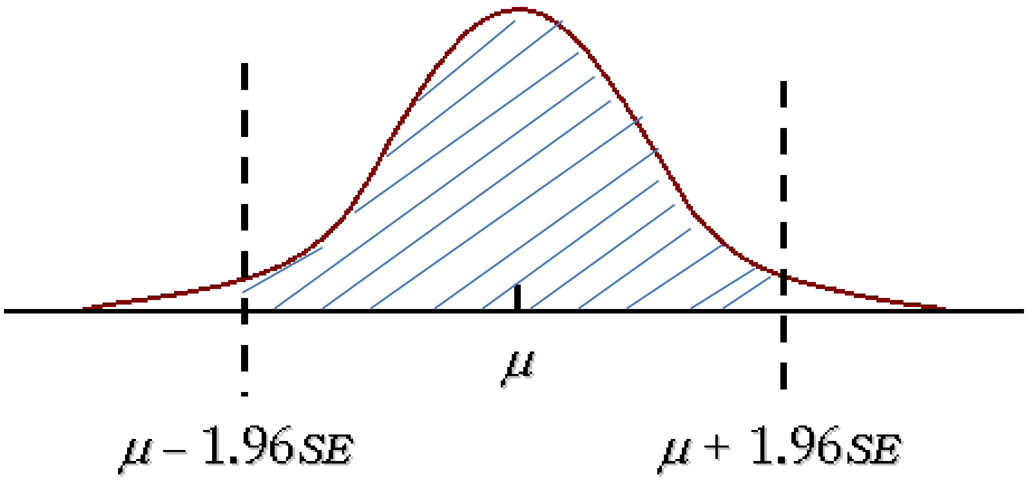
- Based on the sample data:
  - The point estimate for mean balance = $1990
  - Can we trust this estimate?

- What do you think will happen if we took another random sample of 140 alumni?

- Because of this uncertainty, we want to understand how likely is our point estimate within a certain range from the population parameter

Relationship between sample mean and population mean

- For large enough sample size, the distribution of the sample mean $\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$



$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \le \overline{X} - \mu \le 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95$$

- For 95% of the samples, the error of estimation $\le \frac{1.96\sigma}{\sqrt{n}}$

- For a given sample,
  - ✓ A point estimate for the population mean is $\bar{x}$
  - ✓ we are 95% confident that the actual error of estimation $\le \frac{1.96\sigma}{\sqrt{n}}$

ISB

Interval estimate of μ: Confidence interval

- A little bit of math tells us

$$\left(-1.96\,\frac{\sigma}{\sqrt{n}} \leq \bar{X} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) \equiv (\bar{X} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{X} + 1.96\frac{\sigma}{\sqrt{n}})$$

- Interpretation: For 95% of the samples, the interval $(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}}, \bar{x} + 1.96\frac{\sigma}{\sqrt{n}})$ calculated using the sample mean $\bar{x}$ contains μ

- For a given sample, we are 95% confident that μ is in the interval $\boxed{(\bar{x} - \frac{1.96\sigma}{\sqrt{n}}, \bar{x} + \frac{1.96\sigma}{\sqrt{n}})}$

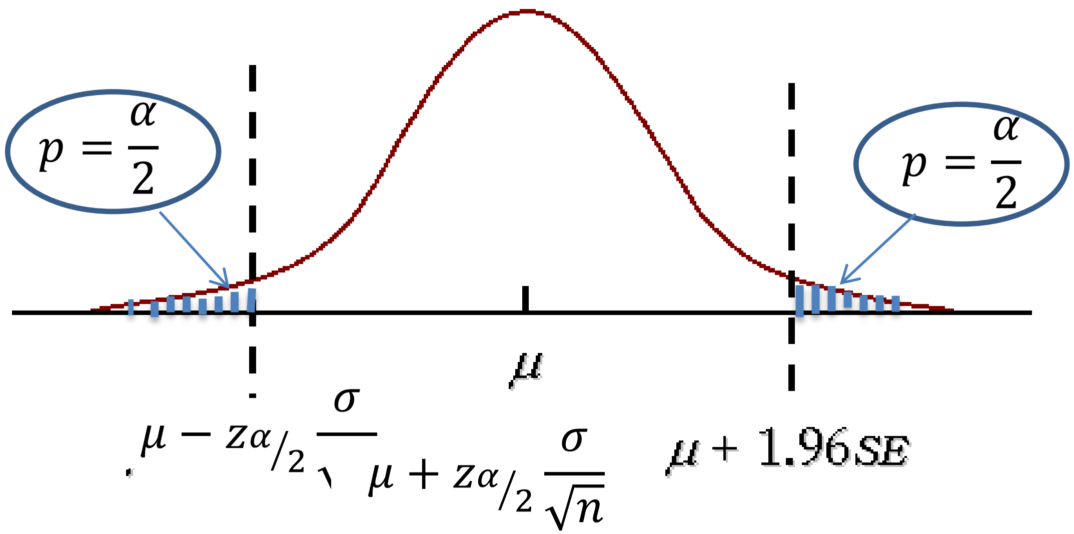
'95% confidence interval for μ'

Interval estimate of μ: Confidence interval

- Start by choosing a confidence level $(1-\alpha)\%$ (e.g. 95%, 99%, 90%)

- Then, we are $(1-\alpha)\%$ confident that the population mean will be within $\left(\bar{x} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}\right)$

is also called ‘margin of error’



Interval Estimate = Point Estimate ± Margin of Error

ISB

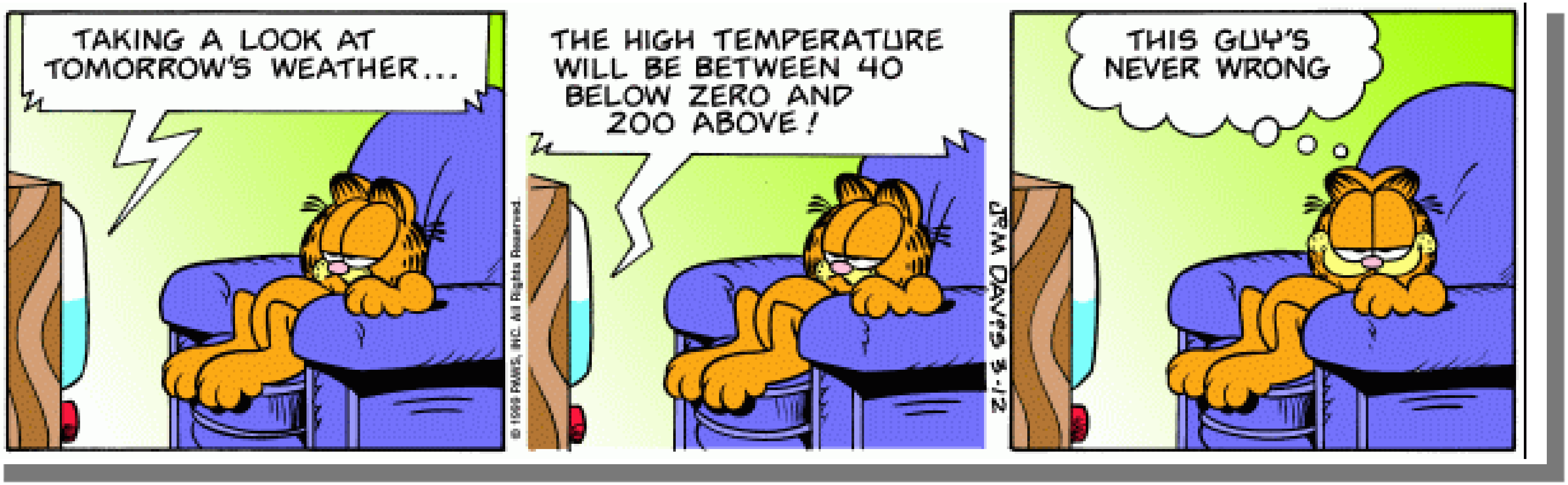
Credit Card: Average balance

- Based on the survey and past data
  - $n = 140, \sigma = 2500, \bar{x} = \$1990$
  - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2500}{\sqrt{140}} = \$211.29$

- Construct a 95% confidence interval for the mean card balance and interpret it

- Does this mean that
  - The mean balance of the population lies in this range?
  - The mean balance is in this range 95% of the time?
  - 95% of the alumni have a balance in this range?

How Big Should be the Margin of Error?



Trade-off between level of confidence and accuracy

- Margin of error depends on the underlying uncertainty, confidence level and the sample size

Credit Card: Average balance

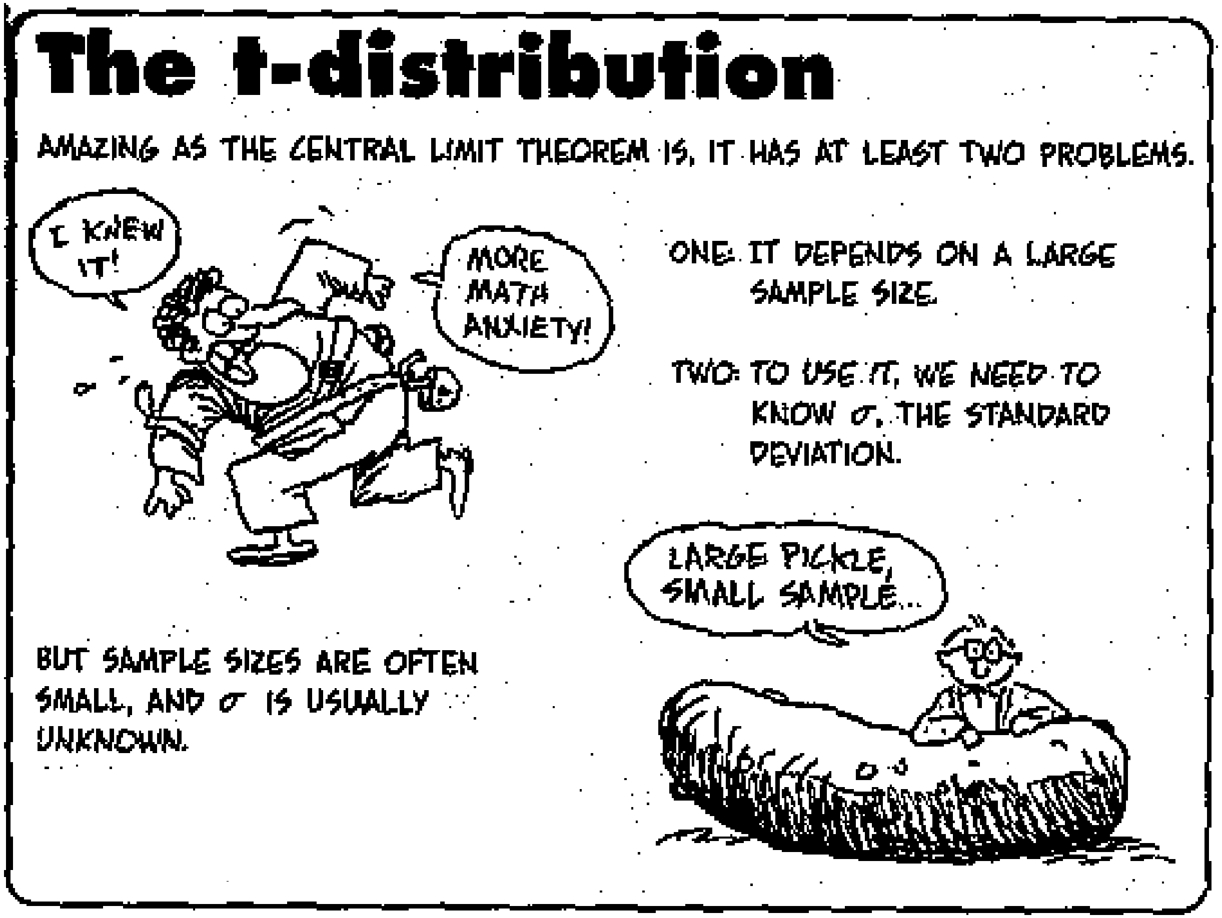
- Based on the survey and past data
  - $n = 140, \sigma = 2500, \bar{x} = \$1990$
  - $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2500}{\sqrt{140}} = \$211.29$

- Construct a 99% confidence interval for the mean card balance and interpret it

- How confident can we be in our estimation if we want the margin of error to be less than or equal to $\$200$?

What if We Don't Know σ ?

- Suppose that the alumni of this university are very different and hence population standard deviation from previous launches cannot be used.



Population Standard deviation Unknown

- We replace σ with our best guess (point estimate) s, which is the standard deviation of the sample:

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

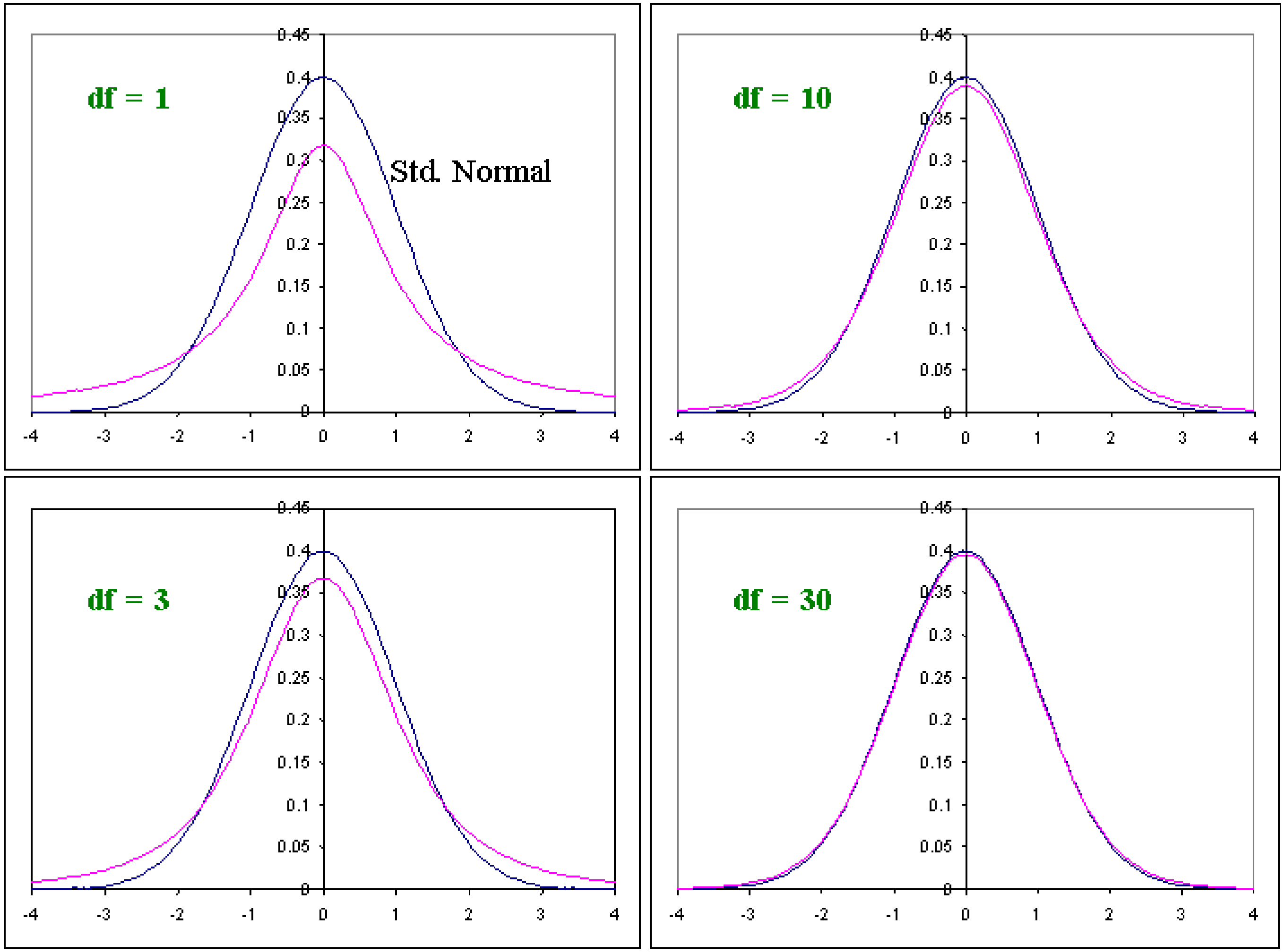
- Let $T = \frac{\bar{X} - \mu}{s/\sqrt{n}}$

- If the underlying population is normally distributed, *T* is a random variable distributed according to a *t*-distribution with *n-1* degrees of freedom (T_{n-1})

- Research has shown that the *t*-distribution is fairly robust to deviations of the population from the normal model

ISB

Student’s T-distribution



As $n \to \infty$,

$t_n \to N(0,1)$

i.e. as the degrees of freedom increase, the *t*-distribution approaches the standard normal dist.

Confidence Interval for Mean with Unknown σ

$$\bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$, where $z_{\alpha/2}$ satisfies $P(Z \geq z_{\alpha/2}) = \alpha/2$

$$\bar{x} \pm t_{\alpha/2,n-1} \frac{s}{\sqrt{n}}$$, where $t_{\alpha/2,n-1}$ satisfies $P(T \geq t_{\alpha/2,n-1}) = \alpha/2$

Calculating t-values

Table entry for p and C is the point t^* with probability p lying above it and probability C lying between $-t^*$ and t^*.

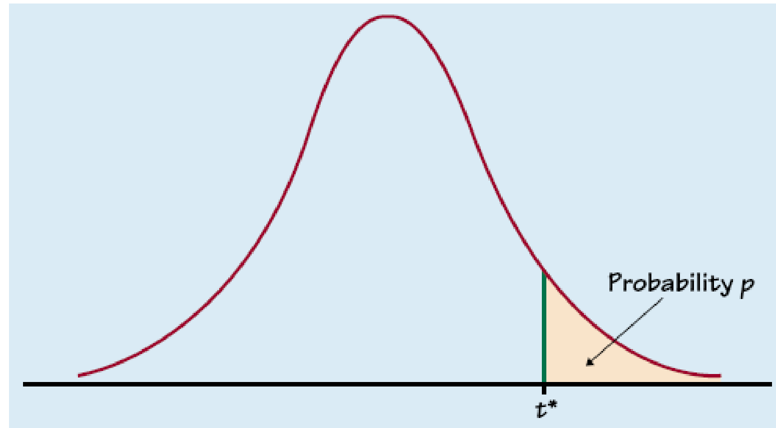


TABLE D ***t* distribution critical values**

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level C											

ISB

Back to credit card balance…

- Recalculate the 95% confidence interval if you cannot assume σ = 2500

- $\frac{\alpha}{2} = 0.025, n = 140$

- Calculate $t_{0.025,139} = 1.98$

- Our estimate of $\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}} = \frac{2833}{\sqrt{140}} = 239.46$

- Then the 95% confidence interval for balance is [$1516, $2464]

ISB

Proportion of Card Acceptance in Population

- Until now, we worked with the sample of alumni who accepted the offer

- However, this represents only 14% of 1000 alumni to whom an offer was sent

- We want to estimate the proportion of the entire population that will accept the card

Distribution of Sample Proportion

- Claim: Proportion is an average of an "appropriately" constructed random variable

- Using CLT, we can establish that sample proportion $P \sim N\left(\pi, \frac{\pi(1-\pi)}{n}\right)$

- The required sample size for CLT to apply
  - n p > 10 and n (1-p) > 10

Confidence Interval for Proportion

- We do not know π, but we use our best guess (point estimate) *p*

- The $(1 - \alpha)\%$ confidence interval can then be specified as $p \pm z_{\alpha/2} \dfrac{\sqrt{p(1-p)}}{\sqrt{n}}$

- What is the 95% confidence interval for the proportion of credit card offers accepted?

How Big a Sample to Get When Estimating Mean?

- Depends on how accurate you want to be, i.e., desired margin of error (DMOE)
  - e.g. Average balance within \$200

- A nutritionist wants to know the average calorie intake for customers to within ± 50 calories with 95% confidence. A pilot study gives an estimate of 430 calories for σ. Find n.

- Actual Margin of error = $t_{\alpha/2,n-1}\dfrac{s}{\sqrt{n}}$

$$n \geq \left(\frac{t_{\alpha/2,n-1}s}{\text{DMOE}}\right)^2$$

- Difficulty: s and n are not known before collecting the sample

- Estimate s from a pilot sample and conduct trial and error to arrive at the appropriate pair of $t_{\alpha/2,\, n-1}$ and n

How Big a Sample to Get When Estimating Proportion?

- Depends on how accurate you want to be, i.e., desired margin of error (DMOE)
  - e.g. Proportion of acceptances within 3%

- What is the sample size required to estimate the proportion of card acceptance within 3% of the population estimate at 95% confidence?

- Actual Margin of error = $z_{\alpha/2}\sqrt{\dfrac{p(1-p)}{n}}$

$$n \geq \left(\frac{z_{\alpha/2}}{\text{DMOE}}\right)^2 p(1-p)$$

- Difficulty: p is not known before collecting the sample

- Utilize the fact that 0≤p≤1 and obtain a conservative estimate with p = 0.5 since it yields the largest value for p(1-p)

ISB

Examples

- A nutritionist wants to know the average calorie intake for female customers to within ± 50 calories with 95% confidence. A pilot study gives an estimate of 430 calories for σ. Find n.
  - Start with z=1.96
  - Calculate $n \geq (1.96 * 430/50)^2 = 284.125 \cong 285$
  - The actual margin of error for this sample size is approximately 50.13, which is greater than DMOE
  - You can fine tune further by increasing ***n*** until you get actual margin of error as exactly 50
  - This is roughly 287

- What is the sample size required to estimate the proportion of card acceptance within 3% of the population estimate at 95% confidence?
  - $n \geq \left(\frac{1.96}{0.03}\right)^2 \frac{1}{4} \cong 1068$

ISB

Summary of Session IV-V

- The sample statistic provides a point estimate
- The interval estimate can be specified by adding and subtracting a margin of error to the point estimate
- The size of the margin of error depends on the level of confidence, the variation in the data and the sample size
- We can use sample standard deviation as an estimate of population standard deviation and use t-values instead of z-values to construct the intervals
- The sample size depends on the margin of error, the standard error and the desired confidence level. For proportions, we can get a conservative sample size by using $p = 0.5$.