Introductory Statistics – Day 22

Intro to Numerical Data

100 Calorie Snack Packs

In recent years, the 100 calorie snack pack has become a popular phenomenon. The packages claim that the snack contains 100 calories. However, we know that we should expect some level of variability in the actual calories in each pack. If a pack contained 105 calories or 92 calories, that would not be terribly unexpected. However, if the **average** calorie count was significantly off from 100, that would be a false advertising problem.

A consumer advocacy group has decided to test the 100 calorie pack claim for a local cookie manufacturer. They suspect that there really are more than 100 calories per pack.

Null Hypothesis: $H_0 : \mu_{calories} = 100$
Alternate Hypothesis: $H_A : \mu_{calories} > 100$

- How might the consumer advocacy group test the advertised claim that the average number of calories is 100? Design a reasonable study.
- If the cookie company is telling the truth,what will the population of cookie packs look like?

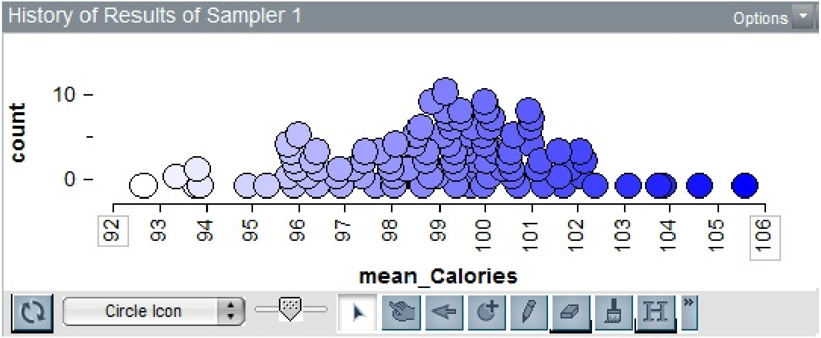
Download the 100CalorieSnackPackSimulator TinkerPlots file. Collect a random sample of 25 snack packs and find the average $\bar{x}_{calorie}$.

- Share your sample mean $\bar{x}_{calorie}$ with the class. Did everyone get the same sample mean? If so, why? If not, how much variability do you see between sample means?
- The consumer advocacy group took a random sample of 25 snack packs and found that the average calorie count was $\bar{x}_{calorie} = 105$ calories.
- How far is this sample mean from the hypothesized 100 calories? (In addition to a numeric answer, indicate whether you think that's a large difference or a small difference.)
- If we want to conduct a hypothesis test, what information do we need to know?
  - Standard Error
  - How far is the sample mean from the expected mean, when measured in standardized units?

$$test\ statistic = \frac{\bar{x} - \mu}{SE}$$

  - How rare is it to find data as extreme as the sample data? (i.e. What's the p-value?)

If 100 students each collected a sample of size 25, we could plot all of these results to see a sampling distribution like the one below. What do you notice about the shape of the distribution?



It's bell shaped!

Three big ideas:

- Point estimates from a sample are useful for estimating population parameters
- Point estimates are not exact. We expect them to vary between samples.
- We can quantify the variability of point estimates. The Central Limit Theorem describes how.

Definition (Central Limit Theorem)

For a population with mean μ and population standard deviation σ, the sampling distribution of the mean approaches a normal distribution. Moreover, we also know what the mean and standard error of this sampling distribution will be.

$$\text{Mean of the sampling distribution} = \mu$$
$$\text{Standard error of the sampling distribution} = \text{SE} = \frac{\sigma}{\sqrt{n}}$$

Fine print - Check the following conditions before assuming that the sampling distribution is nearly normal

- The sample data must be independent.
- The underlying population distribution is not strongly skewed.
- The sample size is large enough. Often $n \geq 30$ is the guideline, but sometimes you can get away with lower, if your underlying population is nicely behaved.

Practicing with the Central Limit Theorem

An IQ test is designed to have a mean of $\mu = 100$ and std. deviation of $\sigma = 15$.

Use the `NORM.DIST()` and `NORM.INV()` commands in Excel to answer the following.

How rare is it for a randomly selected person to have a score of 95 or lower?

How rare is it for a randomly selected group of 30 people to have an average score of 95 or lower?

How rare is it for a randomly selected group of 100 people to have an average score of 95 or lower?

How rare is it for a randomly selected person to have a score of 110 or higher?

How rare is it for a randomly selected group of 30 people to have an average score of 110 or higher?

How rare is it for a randomly selected group of 100 people to have an average score of 110 or higher?

Create the interval that contains the middle 95% of individual test takers.

Create the interval that contains the middle 95% of means for randomly selected groups of 30.

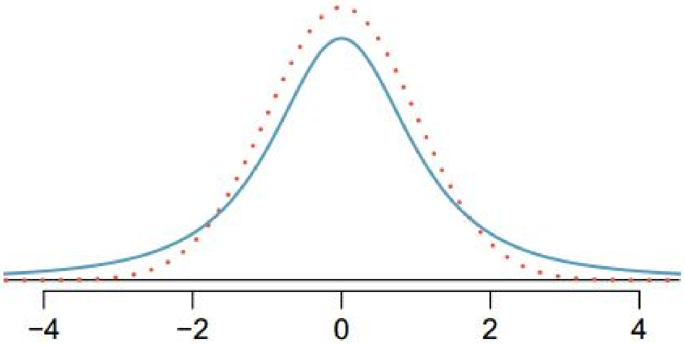
Create the interval that contains the middle 95% of means for randomly selected groups of 100.

In a small town, a random selection of 100 citizens are given the IQ test. Their average score is 108. Is that unusual?

In most real world situations we do not know μ or σ!

Definition (Central Limit Theorem - Modified)

If we do not have μ and/or σ, then the sampling distribution takes on slightly different shape. The sampling distribution is a t-distribution with mean μ and $SE \approx \frac{s}{\sqrt{n}}$



The blue curve is a t-distribution and the red curve is the standard normal distribution. Image from OpenIntro Ch 4.

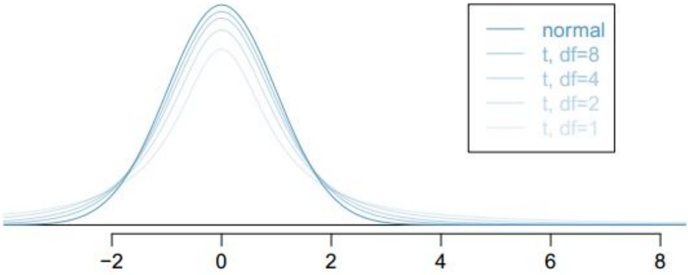


Image from OpenIntro Ch 4.

- t-distribution $\neq$ Normal distribution. How is it similar? How is it different?
- t-distribution $=$ family of mound shaped distributions
- Exact curve depends on the degrees of freedom of the scenario (degrees of freedom is related to sample size).
- As the sample size increases, t-distribution approaches normal distribution.

New vocab term: **Degrees of freedom (df)**

- Sample of size n for one mean problem, $df = n - 1$.
- df determines the shape of the t-distribution.
- larger df, $\Rightarrow$ closer to normal distribution.
- Working with the t-distribution is very similar to working with the normal distribution.
- Excel commands are `T.DIST(test stat, df,1)` and `T.INV(probability, df)`

Practicing with the t-distribution

In order to find the proportion of data to the left of the test statistic x, use the command

```
=T.DIST(test statistic, df, 1).
```

DF refers to degrees of freedom and the final 1 refers to cumulative. The test statistic is computed $\frac{\bar{x} - \mu}{SE}$, the number of standard errors the sample data is from the hypothesized mean.

The command

```
=T.INV(probability, df)
```

is the inverse of T.DIST. A probability between 0 and 1 is entered as the first argument and the output of the funciton is the t-statistic which corresponds with that probability to the left.

Excel commands

In order to find the area to the left of -2.50, with df $=$ 12, use Excel command

```
=T.DIST(-2.5,12,1)
```

In order to find the top 5% of data with df $=$ 12, use the Excel command

```
=T.INV(0.95,12)
```

0.95 is used because that is the proportion of data below our desired cutoff. Remember, always measure from the left side.

- Find the portion of the area to the left of -1.30, with df $=$ 12.
- Find the portion of the area to the left of -1.30, with df $=$ 24.
- Find the portion of the area to the right of 2.30, with df $=$ 10.
- Find the cutoffs for the middle 80% of data, with df $=$ 15.
- Find the cutoffs for the middle 95% of data, with df $=$ 20.