Sampling Distributions, Hypothesis Tests and Confidence Intervals

Dr Tom Ilvento
Department of Food and Resource Economics
University of Delaware

Overview

- We will continue the discussion on Sampling Distributions
- A major theorem is the Central Limit Theorem
- I will introduce how we use this information in a Hypothesis Test
- And a Confidence Interval

We use two theorems to help us make inferences

- In the case of the mean, we use two theorems concerning the normal distribution that help us make inferences
- One depends upon the variable being normally distributed
- The other does not - **Central Limit Theorem**

For Variables that are Distributed Normally

- If repeated samples of a variable Y of size n are drawn from a normal distribution, with mean μ and variance σ^2,
- the sampling distribution of the mean will be a normal distribution with
  - mean μ
  - variance σ^2/n.

$$\frac{\sigma}{\sqrt{n}}$$

For variables that are Distributed Normally

- What we are saying:
  - If we could repeatedly take random samples of size n from a normal distribution.
  - And then take the mean of each sample
  - We would expect the mean of the sample means to equal **μ**
  - And the variance of the sample means would equal **σ^2/n**

Central Limit Theorem

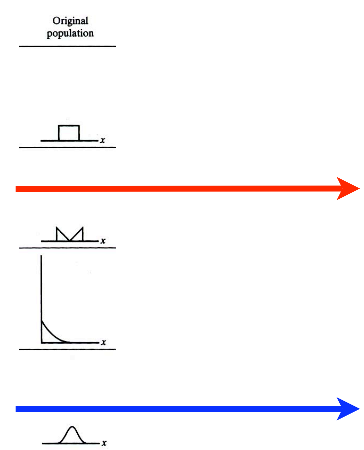
- If repeated sample of Y of size n are drawn from any population (regardless of its distribution as normal or otherwise) having a mean **μ** and variance **σ^2**,
- the sampling distribution of the sample means approaches normality, with **μ** and variance **σ^2/n**.
- **As long as the sample size is sufficiently large**

What is a LARGE n? ~30

Central Limit Theorem

- The Central Limit Theorem is a very powerful for our use.
- It relaxes the assumption of the distribution of the population variable
- *Note: this is based on the notion that our samples are drawn on a **random** probability basis. That is, each element of the population has an equal or near equal chance of being selected*

Demonstration of the Central Limit Theorem



- This shows various distributions of the population
- Next we see a sampling distribution for n=2, small sample
- Then n=5
- Finally, n=30

Inferences from a Sample – my table

- Comparison of the Characteristics of the Population, Sample, and the Sampling Distribution for the Mean

	Population	Sample	Sampling Distribution
Referred to as:	Parameters	Sample Statistics	Statistics
How it is Viewed	Real but not observed	Observed	Theoretical
Mean	$\mu = \frac{\sum X}{N}$	$\bar{x} = \frac{\sum x}{n}$	$\mu = \frac{\sum \bar{x}_n}{\infty}$
Variance	$\sigma^2 = \frac{\sum (X-\mu)^2}{N}$	$s^2 = \frac{\sum (x-\bar{x})^2}{n-1}$	$\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$
Std Deviation	σ	s	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$

Note: I use X and N for the population, and x and n for the sample

How do we use this information?

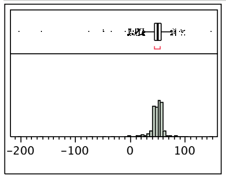
- We draw a random sample
- We think of our sample as one of many possible samples of size n from a population with parameters μ and σ.
- **If the variable is distributed normally**, we can use information about the sampling distribution of the mean to make inferences from the sample to the population.
- Even **if the variable is not distributed normally**, if our sample size (n) is large enough, we can assume the sampling distribution of sample mean is distributed normally (**Central Limit Theorem**)

Let's be detectives and determine is a fraud has been committed!

- A wholesale furniture company had a fire in its warehouse. After determining that the fire was an accident, the company sought to recover costs by making a claim to the insurance company. The company had to submit data to estimate the Gross Profit Factor (GPF).
  - GPF= Profit/Selling Price* 100
- The company **estimated** the GPF based on what was expressed as **a random sample of 253** items sold in the past year and calculated GPF as **50.8%**.
- The insurance company was suspicious of this value and expected a value closer to **48%** based on past experience.
- The insurance company hired us to record all sales in the past year (n=3,005) to calculate a **population GPF**.

Here are the results from the population (n=3005) from JMP

GPF



Quantiles

100.0%	maximum	148.0
99.5%		83.3
97.5%		69.3
90.0%		58.9
75.0%	quartile	56.1
50.0%	median	50.2
25.0%	quartile	44.4
10.0%		38.3
2.5%		16.7
0.5%		-1.8
0.0%	minimum	-202.5

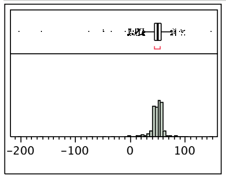
Moments

Mean	48.900978
Std Dev	13.829104
Std Err Mean	0.2522736
upper 95% Mean	49.395625
lower 95% Mean	48.406332
N	3005
Sum Wgt	3005
Sum	146947.44
Variance	191.24412
Skewness	-4.216341
Kurtosis	61.060754
CV	28.279811
N Missing	0

- Let's calculate the z-score and probability for the mean level obtained by the store, 50.8:
- **z = (50.8-48.9)/13.83 = 1.9/13.83 = .137 Small!**
- But the store indicated they took a random sample of 253 items, **so we should use the Std Error for our z-score, based on n=253.**
- **Std Error = 13.829104/SQRT(253) = .8694283**

Now we ask, what is the probability of taking a sample of 253 and getting a sample mean of 50.8, when the true population mean is 48.9?

GPF



Quantiles

100.0%	maximum	148.0
99.5%		83.3
97.5%		69.3
90.0%		58.9
75.0%	quartile	56.1
50.0%	median	50.2
25.0%	quartile	44.4
10.0%		38.3
2.5%		16.7
0.5%		-1.8
0.0%	minimum	-202.5

Moments

Mean	48.900978
Std Dev	13.829104
Std Err Mean	0.2522736
upper 95% Mean	49.395625
lower 95% Mean	48.406332
N	3005
Sum Wgt	3005
Sum	146947.44
Variance	191.24412
Skewness	-4.216341
Kurtosis	61.060754
CV	28.279811
N Missing	0

- Let's re-calculate the z-score and probability for the mean level obtained by the store, 50.8, using a standard error:
- **z = (50.8-48.9)/.869 = 1.9/.869 = 2.186 MUCH LARGER!**
- z = 2.19 is associated with a table value of .4857
- And the probability after that value (2.19) is **.5 - .4857 = .0143**
- **There is something suspicious about the furniture company claim of a mean GPF = 50.8**

Rare Event Approach

- Most inferences will be made using a **Rare Event** approach
  - We will take a sample
  - And compare it to a **hypothesized** population
  - And see how close or far away our sample estimate is from the perspective of a sampling distribution
- We ask, what is the probability of a taking a random sample and observing the sample mean if the population mean is really the hypothesized value
- Or, in the case of a **confidence interval**, we place an interval around our sample estimate using a probability framework

Auto Batteries Example

- The manufacturer claims that life of his automobile batteries is 54 months on average, with a standard deviation of 6 months.
- We are involved in a consumer group and we decide to take a sample of 100 batteries and test the claim.
  - We **select 100 batteries at random**
  - Test them over time and **record the battery life length**
- The mean battery life for our sample is:
  - **Mean = 52 months**
  - **Std Dev = 4.5 months**

Our batteries didn't last as long on average as the manufacturer said, but it is just a sample. How can we test to see if the claim is bogus?

Auto Batteries Example Solution

- If the world works as the manufacturer says
- And I would have taken repeated random samples of size 100
- The **sampling distribution** would be a normal distribution
  - And have a mean equal to the population mean for battery life, i.e., **μ = 54** months
  - And a standard deviation of **σ** divided by the square root of **n**

$$\sigma_{\bar{x}} = \frac{6}{\sqrt{100}} = .60$$

Auto Batteries Example Testing Strategy

- We want to look at our sample as being part of the theoretical Sampling Distribution (SD). That is,
  - SD ~ N (μ, σ/SQRT(n))
  - In this case, **SD ~ N (54, 0.6)**
- And see how likely it is that our sample came from that distribution
- In other words, **how likely is it to get a sample mean of 52 from a random sample of 100 batteries when the true population mean is 54 months?**
- And we will use a z-score and the normal table to help get an answer

How do I do this?

- I hypothesize that the true mean is 54 — **Ho: μ = 54**
- I calculate a z-score based on my sample value (52.0) and the hypothesized mean and standard error (of the sampling distribution)

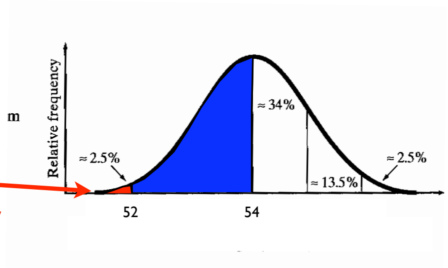
$$z = \frac{(52-54)}{.6} = -3.33$$

- I look up the probability of finding a z-score equal to or less than our calculated value

The p(z=3.33) = not on the table
Using Excel, p = .0004

Answer: Draw It Out!

- z = -3.33 corresponds to a probability of .4996 up to that point
- But I want the point after to get a probability of my "test Statistic"
  - .5 - .4996 = .0004
- **This is a very small probability - a rare event!**
- **This is really a rare event given the claim of the manufacturer – that the batteries really last 54 months on average**



What if we used a sample size = 30?

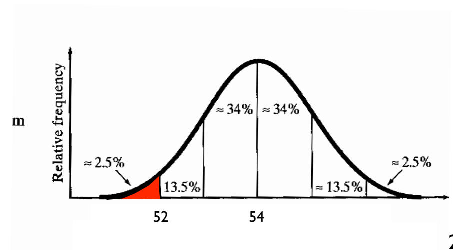
- The standard error of the sampling distribution would change

$$\sigma_{\bar{x}} = \frac{6}{\sqrt{30}} = 1.0954$$

- And the z score would be less out in the tail.

$$z = \frac{(52-54)}{1.0954} = -1.83$$

- Which means the p-value in the tail is now larger
  - p(z=-1.83)= .4664
  - p = .5 - .4664 = .0336
- It still is a rare event, just not as rare



An important point about statistical significance and substantive importance

- Was 2 months on average a big deal?
- Statistics will never tell you the answer on this.
  - The role of statistics is to make an estimate from a sample
  - and give you insight on how rare an event this is based on a probability framework - **statistical significance**
  - It is not about **certainty** or **importance**
- Importance, or **substantive significance**, is based on the discipline

Hypothesis Test

- The battery example is called a **hypothesis test**
- We will develop a more formal approach to a hypothesis test in future lectures
- But the logic is the same
  - We set up a **Null Hypothesis** based on an expectation of nothing happening or a past claim
  - Think of our sample as part of a **sampling distribution**
  - And then see how **rare an event** it was that we drew a sample of size n and achieved a different result from the Null Hypothesis

Confidence Intervals

- We also set up a Confidence Interval
  - We place a **Bound of Error** around our estimate
  - Based on sampling theory
- Confidence Interval involves:
  - The mean
  - plus or minus an interval of the standard error
  - Based on the sampling distribution
  - And a level of probability we are willing to accept

Catalog Sales Data

- I will show output from Excel (Descriptive Statistics) and JMP (Distribution)
- In both cases, we automatically get the **Standard Error** of the Mean
  - Given as **s/SQRT(n)**
- And we can ask for a **Confidence Interval** at a particular **probability level**

	SALES
Mean	1216.77
Standard Error	**30.39**
Median	961.81
Mode	#N/A
Standard Deviation	961.08
Sample Variance	923665.86
Kurtosis	2.97
Skewness	1.47
Coefficient of Variation	78.99
Range	6179.54
Minimum	37.81
Maximum	6217.34
Sum	1216767.86
Count	1000
Confidence Level(95.0%)	**59.64**

95% Confidence Interval (CI) for the Sales data

- The confidence interval gives a plus or minus bound around our estimate
- It is an estimate based on a sample of 1000 customers
- Every estimate comes with some error
- If the only error is sampling error, meaning our estimate is without bias or measurement error, we know what this error should look like in repeated samples

$$\$1216.77 \pm 1.96(30.39) = \$1216.77 \pm 59.56$$

$$\$1{,}157.21 \text{ to } \$1{,}276.33$$

95% Confidence Interval (CI) for the Sales data

$$\$1216.77 \pm 1.96(30.39) = \$1216.77 \pm 59.56$$

$$\$1{,}157.21 \text{ to } \$1{,}276.33$$

- Our **estimate** — $\$1216.77$
- Plus or minus **1.96 standard deviations** — $\$1216.77 \pm 1.96(30.39)$
  - 1.96 in the Standard Normal Table represents a probability of .475
  - 2 * .475 = .95 of a 95% interval
  - Later we will use a t-value of 1.9623
- Where the standard deviation is our **estimate of the standard error**
- The plus/minus part is called a **Bound of Error (BOE)** — $\$1216.77 \pm 59.56$
- The CI is based on the sampling distribution
- We are saying 95% of the intervals constructed this way, based on n = 1000, will contain the population parameter — $\$1{,}157.21 \text{ to } \$1{,}276.33$

Summary

- Sampling distributions are based on repeated samples of an estimator of the same sample size n.
- It gives us a way to begin to make inferences from our sample to the population
- Inferences will come via
  - A z-score based on a hypothesized population parameter
  - A confidence interval of plus or minus so many standard deviations