ØAMET2200
Business Decision Making Using Data
Lecture 2

Instructor: Fenella Carpena

August 30, 2019

Announcements

- Problem Set # 1 has been posted on Canvas
  - Due date: September 5 at 11:59PM
  - You are **encouraged** to work in groups (up to 4 people)
- Office Hours
  - Tuesdays, 4:30-5:30PM and Fridays, 4:30-5:30PM
  - Starting Week 36 (next week)

Part 1 of this Course: Data You Want vs. Data You Have

- Business decision-makers never have all the data they want
  - about their customers
  - about their employees' productivity
  - about their competitors
- If you can only collect **a sample of data**
  - How should you select that sample?
  - What can you say about the overall "population"?

Agenda for Today

- Samples and Surveys (Chapter 13)
  - Sampling methods
  - Biased samples
- Sampling Variation and Quality (Chapter 14.1-14.2)
  - Sampling distribution of the mean
  - Control limits
- Stata
  - `do` and `log` files
  - Exercise

J.D. Power and Associates

Dinside Motor Teknologi Økonomi Bolig Tester Kalkulatorer Snarveier Søk

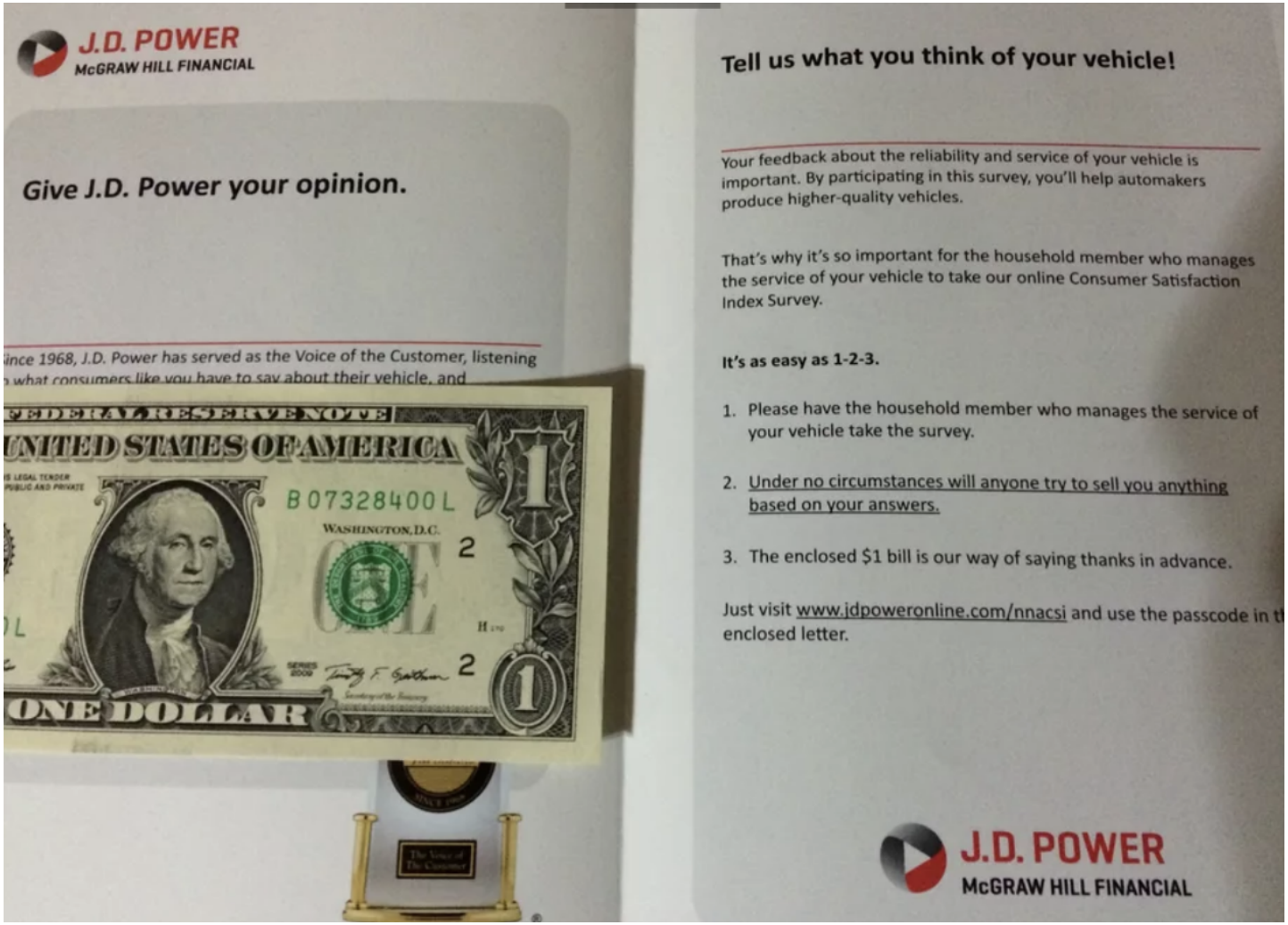
J. D. Power og bilkvalitet 2019

Kia og Hyundai med færrest feil - igjen

Sørkoreanske bilmerker befester posisjonen med færrest feil på markedet.

- A global marketing information services firm, specializing in customer satisfaction
- Best known product is the "Initial Quality Study," which assesses car quality
- How is the best car model determined?

How is the best car model determined?

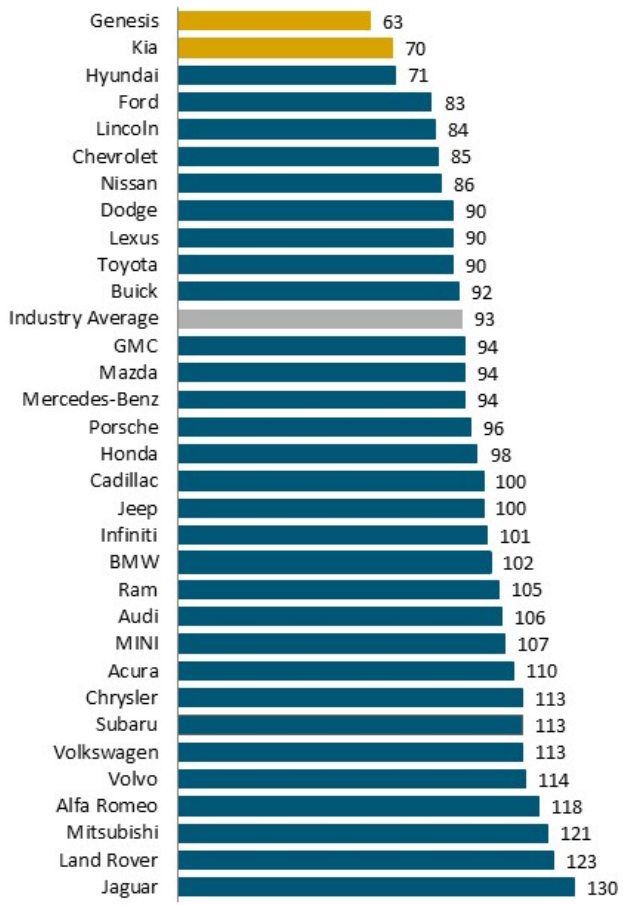


- Focus on a subset of new-car owners (a sample)
- Should ensure that survey respondents are selected randomly from the population

Initial Quality Study, 2019 Results

J.D. Power
2019 U.S. Initial Quality Study℠ (IQS)

2019 Brand Ranking
Problems per 100 Vehicles (PP100)



Note: Included in the study, but not ranked due to small sample size, is Fiat. Not included in the ranking due to unrepresentative sample size is Tesla.

Source: J.D. Power 2019 U.S. Initial Quality Study℠ (IQS)

Definitions

- Population: the entire collection of interest
- Sample: a subset of the population
- Survey: ask questions to a sample to learn about the population
- Representative: samples that reflect the mix in the entire population; **this is the type of sample we want**
- Bias: systematic error in selecting the sample

There is **one population** but **many different samples**.

How do we get a representative sample?

- The best way to get a representative sample is to pick members of the population **at random**
- A **random sample** is representative of the whole population
- Randomization **eliminates bias** and allows us to infer characteristics of the population from the sample
- Randomization ensures that, **on average**, a sample mimics the population

Example: Two Random Samples from a Population

	Age (years)	Percentage Female	Number of Children	Income Category	Percentage Homeowners
Sample 1	45.12	31.54	1.91	5.29	61.4
Sample 2	44.44	31.51	1.88	5.33	61.2
Population	**44.88**	**31.51**	**1.87**	**5.27**	**61.1**

- Suppose we have a database of all individuals who bought a new car in 2018 (3.5 million people)
- We took one sample of 8,000 people (Sample 1)
- We took another sample of 8,000 people (Sample 2)
- Notation: N (population size) and n (sample size)
- What's the point? We can't survey survey the whole population, but **with randomization, we obtain samples that look like the population**

How do we randomize?

- We take a **simple random sample** (SRS)
  - A sample of size n is chosen randomly from the population
  - SRS assigns an equal chance to every *combination* of n members of the population
  - SRS is the gold standard to which all other sampling methods are compared
- To get a SRS, we first need to identify the **sampling frame**
  - A list from which to draw the sample
  - We want the sampling frame to list every member of the population of interest
  - But often, the list we have $\neq$ the list we want
  - Example: Election polling
    - Population of interest: people who **will** vote
    - Sampling frame: people aged 18 and above who **can** vote

Algorithm for taking a SRS

Suppose we have a database of individuals who bought a new car in 2018, and we want a sample size of $n = 5$ people for the survey

Step 1: Open the database (this is our sampling frame)

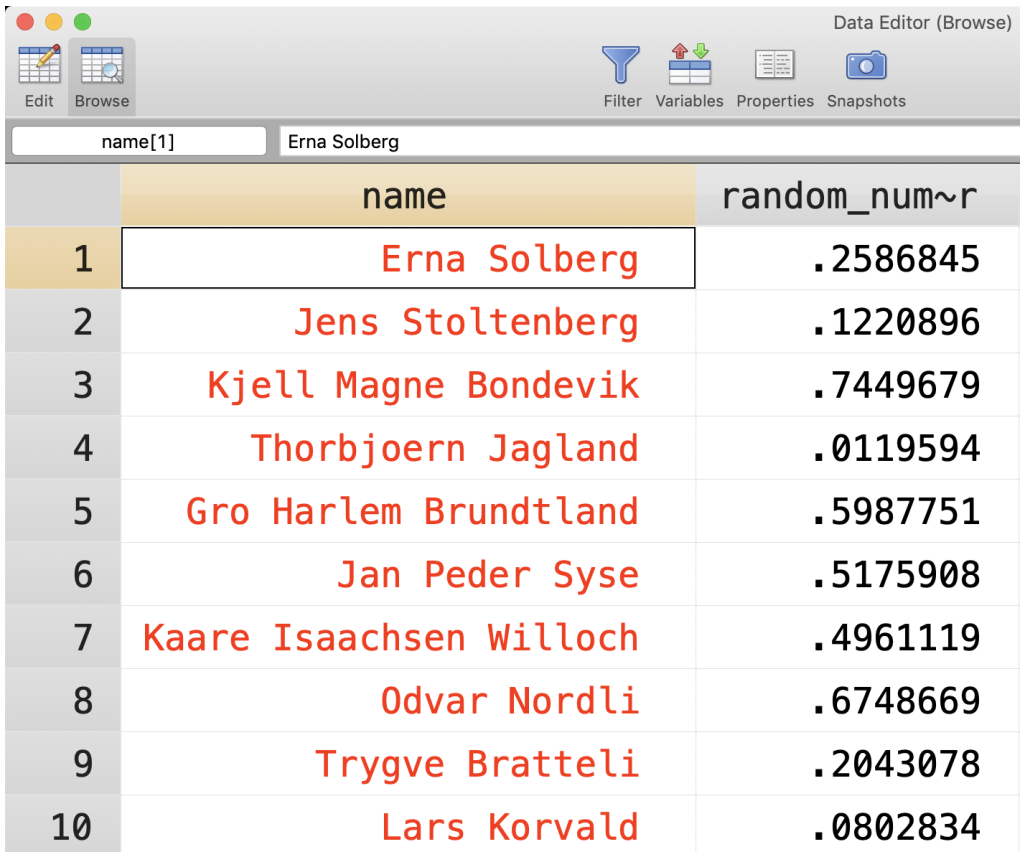
In Stata: `use /users/fenella/desktop/sampling_frame.dta`

	name
1	Erna Solberg
2	Jens Stoltenberg
3	Kjell Magne Bondevik
4	Thorbjoern Jagland
5	Gro Harlem Brundtland
6	Jan Peder Syse
7	Kaare Isaachsen Willoch
8	Odvar Nordli
9	Trygve Bratteli
10	Lars Korvald

Algorithm for taking a SRS

Step 2: Add a column of random numbers

In Stata: `generate random_number = runiform()`

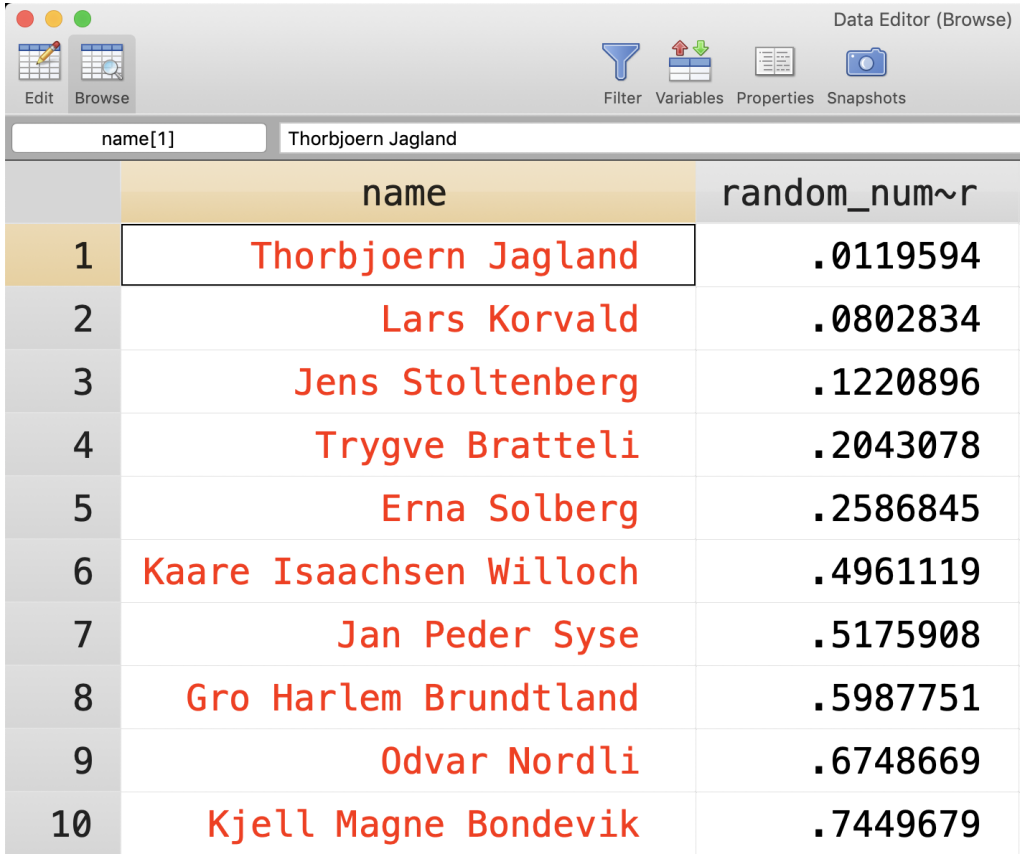


	name	random_num~r
1	Erna Solberg	.2586845
2	Jens Stoltenberg	.1220896
3	Kjell Magne Bondevik	.7449679
4	Thorbjoern Jagland	.0119594
5	Gro Harlem Brundtland	.5987751
6	Jan Peder Syse	.5175908
7	Kaare Isaachsen Willoch	.4961119
8	Odvar Nordli	.6748669
9	Trygve Bratteli	.2043078
10	Lars Korvald	.0802834

Algorithm for taking a SRS

Step 3: Sort the rows using the random numbers from Step 2

In Stata: `sort random_number`



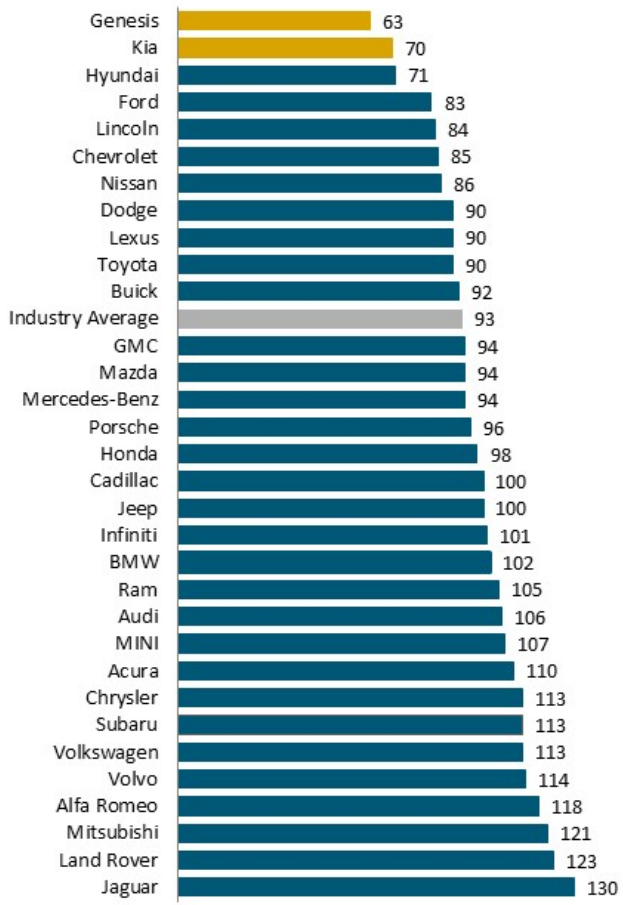
	name	random_num~r
1	Thorbjoern Jagland	.0119594
2	Lars Korvald	.0802834
3	Jens Stoltenberg	.1220896
4	Trygve Bratteli	.2043078
5	Erna Solberg	.2586845
6	Kaare Isaachsen Willoch	.4961119
7	Jan Peder Syse	.5175908
8	Gro Harlem Brundtland	.5987751
9	Odvar Nordli	.6748669
10	Kjell Magne Bondevik	.7449679

Step 4: Those in the first $n = 5$ rows identify a SRS

What are potential problems with a SRS?

J.D. Power
2019 U.S. Initial Quality Study℠ (IQS)

2019 Brand Ranking
Problems per 100 Vehicles (PP100)



Note: Included in the study, but not ranked due to small sample size, is Fiat. Not included in the ranking due to unrepresentative sample size is Tesla.

Source: J.D. Power 2019 U.S. Initial Quality Study℠ (IQS)

Charts and graphs extracted from this press release for use by the media must be accompanied by a statement identifying J.D. Power as the publisher and the study from which it originated as the source. Rankings are based on numerical scores, and not necessarily on statistical significance. No advertising or other promotional use can be made of the information in this release or J.D. Power survey results without the express prior written consent of J.D. Power.

Alternatives to SRS

1. Stratified Random Sample

- A sample derived from random sampling within subsets of similar items known as strata
- Two-step process:
  - Divid the sampling frame into strata
  - Use SRS to select items from each strata
- Why do we care? Stratified random sampling allows us to capture specific portions of the population

Alternatives to SRS

2. Cluster Sampling

- A type of stratified sampling that groups the population into conveniently surveyed geographic clusters
- Three-step process:
  - Divide the population into geographic clusters
  - Randomly select clusters
  - Randomly choose item within selected clusters
- Why do we care? Can reduce the costs of interviewing, especially in an in-person national survey

Alternatives to SRS

3. Voluntary Response Sample

- A sample of people who volunteered to participate in a survey
- Not representative of the population; usually biased towards those with strong opinions

4. Convenience Sampling

- A sampling method that selects those who are readily available
- Not representative of the population
- Example: Surveys done at a shopping mall/on the streets; the interviewer tends to select people who look easy to interview

Biased samples

- We assume we have **unbiased samples** in much of this course
- But you **always** need to think about biased samples
- Two issues to think about:
  1. **Selection** into the sample
  2. Survey **design** and **implementation**

Selection into sample

- **Key question:** Who is in the sample? Who is not?
- Those included in the sample should **not be** systematically different from those excluded
- Some potential sources of bias
  - Response bias: those who respond to the survey are not the same as the average person in the population
  - Survivor bias: "Planes coming home from battle have bullet holes everywhere but the engine and cockpit, so we should put armor everywhere but the engine and cockpit.

Survey design and implementation

- How were questions asked? Survey wording matters
- Example from a Pew Research Survey
  - Should it be illegal for doctors to **give terminally ill patients the means to end their lives?** 51% said yes
  - Should it be illegal for doctors to **assist terminally ill patients in committing suicide?** 44% said yes
- For in-person interviews, the **interviewer** and the **actual interview situation** (e.g., location) can also affect responses

In summary: Some questions to ask whenever you get data

- What was the sampling frame? Does it match the population?
- Is the sample a SRS? If not, what is the sampling design?
- If data are from a survey:
  - What is the rate of nonresponse?
  - How was the question worded?
  - Did the interviewer affect the results?
- Does survivor bias affect the data?

Estimating Parameters

- Using the data from our sample, we can calculate **sample statistic**, a characteristic observed in the sample
- Our goal is to use sample statistics to say something about **population parameters**, a characteristic of the population (usually unobserved)
- Sample statistics are used to **estimate** the population parameters
- For example, we might use the **sample mean** to estimate the **population mean**

Notation: Sample Statistics and Population Parameters

Name	Sample Statistic	Population Parameter
Mean	$\bar{y}$	μ (mu, pronounced "mew")
Standard deviation	s	σ (sigma)
Correlation	r	ρ (rho, pronounced "row")
Slope of line	b	β (beta)
Proportion	$\hat{p}$	p

Sampling Variation

- There is **one population** but **many different samples**
- So the value of the sample statistic will differ from sample to sample; this variability is called **sampling variation**
- It is the price we pay for working with a sample rather than the population
  - For example, the sample mean won't necessarily be close to the population mean
  - But if we have a random sample, statistics allow us to quantify the effects of sampling variation and reach conclusions about the population

Example of Sampling Variation: Coin Toss

Imagine flipping a coin 10 times. If the coin is fair, what is the proportion of heads that you should get?

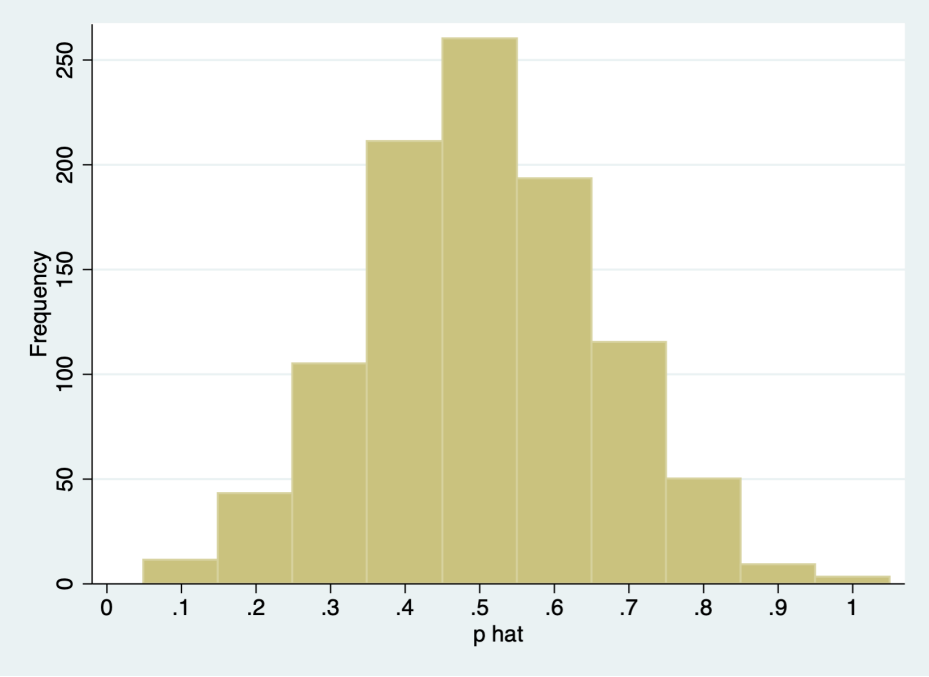
Sample # 1 

Sample # 2 

Sample # 3 

Example of Sampling Variation: Coin Toss

Do this 1,000 times, and we get 1,000 values of $\widehat{p}$. If we plot these $\widehat{p}$ in a histogram, it would look like this:



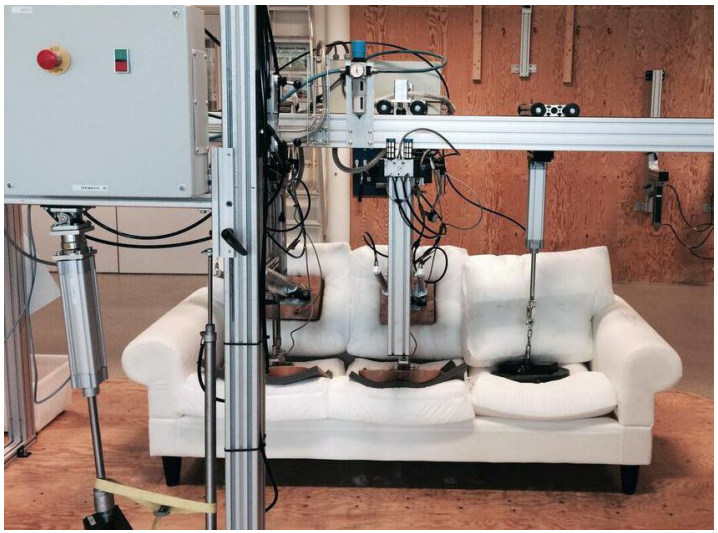
- Most $\widehat{p}$ cluster around $p = 0.5$, the population proportion
- But some $\widehat{p}$ are lower or higher than 0.5
- This variability in the value of $\widehat{p}$ from sample to sample is **sampling variation**

Sample Statistics are Random Variables

- Because of sampling variation, sample statistics (e.g., the sample proportion $\widehat{p}$, the sample average $\widehat{X}$) are **random variables**)
- Sample statistics have a mean (e.g., $E(\widehat{p})$, $E(\overline{X})$) and variance (e.g., $Var(\widehat{p})$, $Var(\overline{X})$)
- Sample statistics have a probability distribution; this distribution is called the **sampling distribution**

Application: Sampling Variation and Quality Control

What is quality control? How do manufacturers test quality?



Application: GPS Chips

- Manufacturers use a Highly Accelerated Life Test (HALT) to check that GPS chips are manufactured as designed

Test No.	Vibration	Voltage	Temperature
1	Low	Normal	50°
2	None	+15%	50°
⋮	⋮	⋮	⋮
15	High	+60%	130°

- Engineers take a random sample of 20 chips everyday
- Each sampled chip undergoes HALT
  - If the chip fails on the 1st test, HALT score = 1
  - If the chip fails on the 2nd test, HALT score = 2
  - If the chip endures all 15 tests, HALT score = 16
  - Each sampled chip will have a HALT score (an integer 1 to 16)

Random Variation or Faulty Production?

- Suppose that engineers who designed the production line say that when manufacturing is functioning properly, on average the chips should have a HALT score of 7 with a SD of 4
  - The population parameters are $\mu = 7$ and $\sigma = 4$
  - Let's also assume that the HALT scores are independent
- Even when manufacturing is functioning properly, there is variation in HALT scores
  - Not every chip is going to have a HALT score of 7: some chips will do better, other chips will do worse
- **How can we use the sample of 20 HALT scores to decide if production is working well (or to stop production)?**
  - Take the sample average of the HALT scores, $\overline{X}$
  - For large n, $\overline{X}$ is approximately normally distributed
  - Use normal distribution to construct **control limits**

The Sample Average $\overline{X}$

- Also called the **sample mean**
- Average of n individual measurements from a population, where n is the sample size

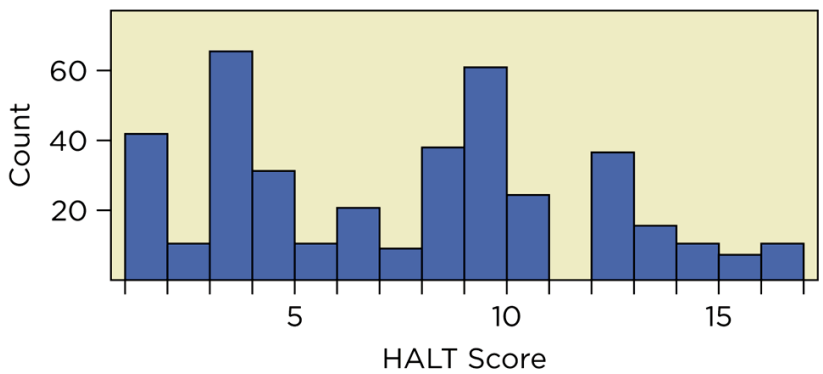
$$\overline{X} = \frac{X_1 + X_2 + \ldots + X_n}{n}$$

- Each X_i is drawn from **the same population**, so all X_i's have the same mean μ and SD σ
- Example: GPS chips
  - We have $n = 20$, so $\overline{X} = \frac{X_1 + X_2 + \ldots + X_{20}}{20}$
  - Each of $X_1, \ldots, X_{20}$ is a HALT score of a sampled chip
  - $X_1, \ldots, X_{20}$ all have the same mean ($\mu = 7$) and SD ($\sigma = 4$)
- **Important!** $\overline{X}$ is a r.v. and it has a **sampling distribution**

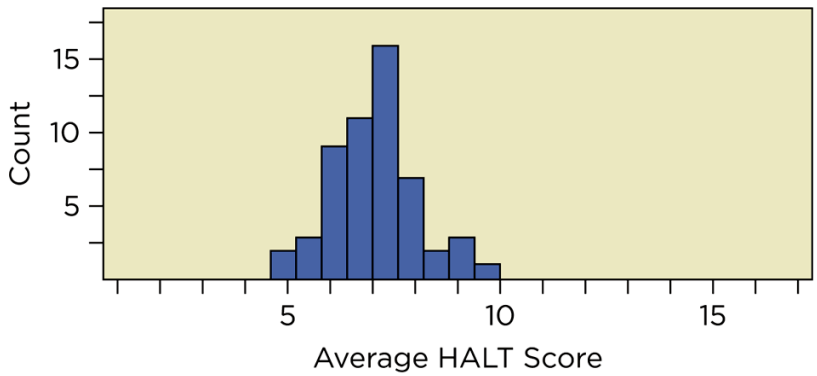
Example: GPS chips

If the manufacturing process is working as designed, what is $E(\overline{X})$? What is $Var(\overline{X})$? Assume that HALT scores are independent.

Benefits of Averaging: (1) Reduces Variance



(a) Individual HALT scores

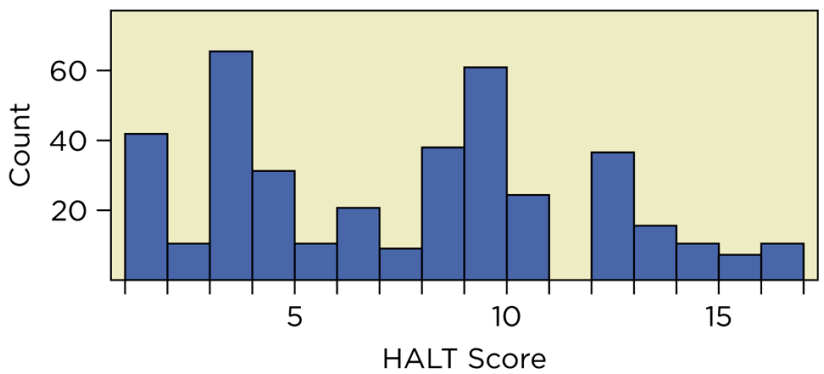


(b) Average HALT scores across 54 days, samples of 20 chips per day

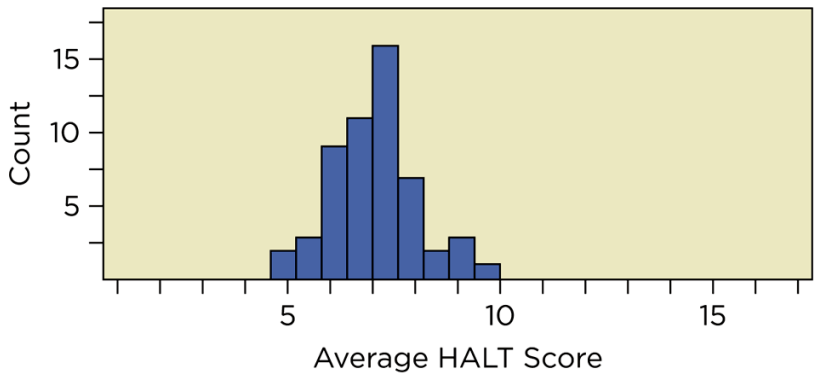
The sample-to-sample variance of daily average HALT score is **smaller** than the variance of individual HALT scores

- If we look at 1 chip at random, we might get a really "good" chip or a really "bad" chip
- If we look at 20 chips, more likely a mix of "good" and "bad" chips; a sample of 20 "bad" chips less likely

Benefits of Averaging: (2) Normality



(a) Individual HALT scores



(b) Average HALT scores across 54 days, samples of 20 chips per day

The distribution of the **sample average** will be approx. **normal**, regardless of the distribution of the underlying observations

- For large enough sample size n, follows from Central Limit Theorem (*sentralgrenseteoremet*)
- Why do we care? Allows us to define "unusual" outcomes, which would signal faulty production

Central Limit Theorem (*sentralgrenseteoremet*)

- Suppose that $E(X_i) = \mu$ and $SD(X_i) = \sigma$. The theorem tells us that for sufficiently large n, $\overline{X}$ is approximately normally distributed with mean μ and variance σ^2/n. That is,

$$\overline{X} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$$

- Example: GPS chips. If the manufacturing process is working as designed,

$$\overline{X} \sim \mathcal{N}\left(7, \frac{4^2}{20}\right)$$

- When is the sample size **large enough**? We use the following **sample size condition**
  - A normal model provides an accurate approximate of the sampling distribution of $\overline{X}$ if the sample size n is larger than 10 times the absolute value of the sample kurtosis, $n > |10 \cdot K_4|$

What is kurtosis?

- Denoted as K_4 and calculated as

$$K_4 = \frac{z_1^4 + z_2^4 + \ldots + z_n^4}{n} - 3, \text{ where } z_i = \frac{x_i - \overline{x}}{s}$$

- If the data are normally distributed, $K_4 \approx 0$
- Measures how much mass is in the tails, and therefore, how much of the variance arises from outliers
- Higher kurtosis means that outliers are more likely
- In our GPS chips example: suppose we find $K_4 \approx -1$.
  - Do our data meet the sample size condition?
  - What model describes the sampling distribution of $\overline{X}$?

Recap

- The **sample mean** is $\overline{X} = \frac{X_1 + X_2 + ... + X_n}{n}$
- Each X_i is drawn from the **same population**, so they all have the same mean μ and SD σ
- $\overline{X}$ is a **random variable** because the value of $\overline{X}$ differs every time we take a different sample
- Because $\overline{X}$ is an r.v., it has a probability distribution function; this function is called the **sampling distribution**
- Even though we may not know what the sampling distribution of $\overline{X}$ is, we can use the **Central Limit Theorem**
  - If the sample size condition is met, $\overline{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- Why do we care about the sampling distribution of $\overline{X}$? It tells us if the sample average we seen in our data is a common outcome
  - In our GPS chips example: an uncommon sample average could mean that production is not working well

Control Limits

- Sampling distribution tells us what the typical outcomes are when the manufacturing process is working as designed
  - We should expect the sample average HALT score to be near 7
- What if the sample average HALT score we get is 5?
  - It could be that the process is operating correctly, but with "bad luck" we got a rare outcome in our sample
  - It could also be that the process is not working as designed
- How to make a decision to continue or to stop production?
- **Control limits** provide an approach to this question

Control Limits

- A **control limit** is a threshold that determine whether a process should be allowed to continue or not

- If $\overline{X}$ lies within the range

$$\mu - L \leq \overline{X} \leq \mu + L$$

  then $\overline{X}$ is close enough to μ and we should continue production; otherwise, $\overline{X}$ is too far from μ and we should stop the process to find the problem

- To construct the control limit, we need L

How to pick L in the control limit $\mu - L \leq \overline{X} \leq \mu + L$?

- The choice of L affects the likelihood of Type I and Type II errors
- Type I error: taking action when you don't need to (false positive)
- Type II error: failing to act when you should have (false negative)

		Supervisor Chooses to	
		Continue	**Shut Down**
State of process	**Working as designed**	✓	$\times_1$
	Not working as designed	$\times_2$	✓

Exercise: Type I or Type II Error?

(a) A jury convicts an innocent defendant

(b) A retailer fails to stock fashion items that become popular in the coming season

(c) A diagnostic test fails to detect the presence of a serious virus infection

(d) A company hires an applicant who is not qualified for the position

Example: GPS Chips

Suppose that the production line is shut down when the sample mean HALT score is less than 6 or more than 8, what is the probability of Type I error?

Balancing Type I and Type II Errors

- The control limit is $\mu - L \leq \overline{X} \leq \mu + L$
- Should we pick a big or small L?
  - A big L (wide control limit) reduces the chance of Type I error
  - A small L (narrow control limit) reduces the chance of Type II
  - Cannot simultaneously reduce the chances for both
- Control limits are determined by focusing on Type I errors
- The probability of a Type I error is typically set at 5% or 1%
  - $P(|Z| > 1.96) = 0.05$
  - $P(|Z| > 2.58) = 0.01$

Example: GPS Chips

Suppose the manager's tolerance for Type I error is 5%. What is the control limit that he/she should set?