Ch1

Resampling

Here's how the process works. On a set of training data:

1. Build a model to predict the target variable (output) and note its strength (e.g., R-squared, lift, correlation, explanatory power).
2. Randomly shuffle the target vector to "break the relationship" between each output and its vector of inputs.
3. Search for a new best model – or "most interesting result" - and save its strength. (It is not necessary to save the model; its details are meaningless by design.)
4. Repeat steps 2 and 3 many times and create a distribution of the strengths of all the bogus "most interesting" models or findings.
5. Evaluate where your actual results (from step 1) stand on (or beyond) this distribution. This is your "significance" measure or probability that a result as strong as your initial model can occur by chance.

Explanation

Let's break this down: imagine you have a math class full of students who are going totake a quiz. Before the quiz, everyone fills out a card with specified personal information,such as name, age, how many siblings they have, and what other math classes they've taken.Everyone then takes the quiz and receives their score.To discover why certain students scored higher than others, you could model the targetvariable (the grade each student received) as a function of the inputs (students' personalinformation) to identify patterns. Let's say you find that older sisters have the highest quizscores, which you think is a solid predictor of which types of future students will performthe best.But depending on the size of the class and the number of questions you asked everyone,there's always a chance that this relationship is not real, and therefore won't hold true forthe next class of students. (Even if the model seems reasonable, and facts and theory can bebrought to support it, the danger of being fooled remains: "Every model finding seems tocause our brains to latch onto corroborating explanations instead of generating the criticalalternative hypotheses we really need.")With target shuffling, you compare the same inputs and outputs against each other asecond time to test the validity of the relationship. This time, however, you randomly shufflethe outputs so each student receives a different quiz score—Suzy gets Bob's, Bob getsEmily's, and so forth.All of the inputs (personal information) remain the same for each person, but each nowhas a different output (test score) assigned to them. This effectively breaks the relationshipbetween the inputs and the outputs without otherwise changing the data.You then repeat this shuffling process over and over (perhaps 1000 times, though even 5times can be very helpful), comparing the inputs with the randomly assigned outputs eachINTRODUCTIONxixtime. While there should be no real relationship between each student's personal informa-tion and these new, randomly assigned test scores, you'll inevitably find some new falsepositives or "bogus" relationships (e.g. older males receive the highest scores, women whoalso took Calculus receive the highest scores, etc.).As you repeat the process, you record these "bogus" results over the course of the 1000random shufflings. You then have a comparison distribution that you can use to assesswhether the result that you observed in reality is truly interesting and impressive or to whatdegree it falls in the category of "might have happened by chance."Elder first came up with target shuffling 20 years ago, when his firm was working witha client who wasn't sure if he wanted to invest more money into a new hedge fund. Whilethe fund had done very well in its first year, it had been a volatile ride, and the client wasunsure if the success was due to luck or skill. A standard statistical test showed that theprobability of the fund being that

successful in a chance model was very low, but the clientwasn't convinced.So Elder performed 1,000 simulations where he shuffled the results (as described above)where the target variable was the daily buy or hold signal for the next day. He then compared the random results to how the hedge fund had actually performed.Out of 1,000 simulations, the random distribution returned better results in just 15instances—in other words, there was only a 1.5% chance that the hedge fund's successcould occur just as the result of luck. This new way of presenting the data made sense tothe client, and as a result he invested 10 times as much in the fund.1"I learned two lessons from that experience," Elder says. "One is that target shuffling is avery good way to test non-traditional statistical problems. But more importantly, it's a pro-cess that makes sense to a decision maker. Statistics is not persuasive to most people—it'sjust too complex."If you're a business person, you want to make decisions based upon things that are realand will hold up. So when you simulate a scenario like this, it quantifies how likely it is thatthe results you observed could have arisen by chance in a way that people can understand."

Data Scientist

Around 2010, the termdata scientistcame into use to describe analysts who combinedthese two sets of skills. Job announcements now carry the termdata scientistwith greaterfrequency than the termstatistician, reflecting the importance that organizations attach tomanaging, manipulating, and obtaining value out of their vast and rapidly growing quanti-ties of data

- Statistician
- Programmer

1.1 It is indeed possible with a probability of 2/12

Try It Yourself 1.1

Let us look first at the idea of randomness via a classroom exercise.

1. Write down a series of 50 random coin flips without actually flipping the coins. That is, write down a series of 50 Hs and Ts selected in such a way that they appear random.
2. Now, actually flip a coin 50 times.

If you are reading this book in a course, please report your results to the class for compilation—specifically, report two lists of Hs and Ts like this: My results—Made up flips: HTHHHTT, and so on. Actual flips: TTHHTHTHTH, and so on.

1.2 Not really, I get zero given that there are too few observations in each trial.

Try It Yourself 1.2

Let us double the sample size and imagine that the study had revealed 14 errors in 1 year and six the following, instead of seven and three. Now, regroup your 12 simulations of 10 tosses into six trials of 20 tosses each. Combine the first and the second sets, the third and fourth, and so on. Then do six more trials of 20 tosses each for a total of 120 additional tosses. You should now have 12 sets of 20 tosses. If you want to try a computer simulation, use the Box Sampler macro-enabled Excel workbook boxsampler2.xlsm.

The textbook supplements contain a Resampling Stats procedure for this problem. Did you ever get 14 or more heads in a trial of 20 tosses?

Observational study: being able to go over historical data, look for patterns, and draw conclusions.
Experiment: collecting data to answer a pre-defined question

This is a personal research, not related to the book :)
N-1 vs n

Formula [edit]

The sample mean is given by

$$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

The biased sample variance is then written:

$$s_n^2 = \frac{1}{n}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} \left(x_i^2\right)}{n} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{n^2}$$

and the unbiased sample variance is written:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n} (x_i - \bar{x})^2 = \frac{\sum_{i=1}^{n} \left(x_i^2\right)}{n-1} - \frac{\left(\sum_{i=1}^{n} x_i\right)^2}{(n-1)n} = \left(\frac{n}{n-1}\right) s_n^2.$$

Check for proof

standard measures with which statisticians are concerned: **central location** of and **variation in the data**.

Central Location

X-bar → sample mean

Mu → population mean

Median→ is concerned with the center of the data, and misses what is going on with the beginning and end of the data (a typical outcome).

Mode→ most common data point (most common sold car in usa Buick…)

Expected Value → Multiply each outcome by its probability of occurring.2. Sum these valueThe expected value is really a fancier mean; it adds the ideas of future expectations and probability weights.

Question 1.3A student gave seven as the median of the numbers 3, 9, 7, 4, 5. What do you think he or she did wrong?

-Wrong, median looks at sorted data :)

VARIABILITY

Variability lies at the heart of statistics: measuring it, reducing it, distinguishing ran-dom from real variability, identifying the various

sources of real variability, and making decisions in the presence of it.Just as there are different ways to measure central tendency—mean, median,mode—there are also different ways to measure variability.

Range→ max-min, sensitive to outliers
Percentile → In a population or a sample,

Def:the Pth percentile is a value such that at least P percent of the values take on this value or less and at least (100−P) percent of the values take on this value or more.

Intuitively: to find the 80th percentile

- sort the data.
- Then, starting with the smallest value, proceed 80% of the way to the largest value

IQR → Q3 -Q1 75% -25% (also called *hinges*)

Residuals → deviations from some typical value

Def:A residual is the difference between a mean value and an observed value or the difference between a value predicted by a statistical model and an actual observed value.

Mean Absolute Deviation → Typical value for residuals

Def: We could take the absolute values of the deviation and average them. Taking the deviations themselves, without taking the absolute values, would not tell us much—the negative deviations exactly offset the positive ones. This always happens with the mean.

Variance → Another way to deal with the problem of positive residuals offsetting negative ones is by squaring the residuals.

Def:Variance for a population is the mean of the squared residuals, whereμ=population mean,x represents the individual population values, and N=population size

$$\text{Variance} = \sigma^2 = \frac{\sum (x - \mu)^2}{N}$$

Standard deviation → $\sigma$is the square root of the variance. The symbol σ is the Greek letter sigma and commonly denotes the standard deviation.

The standard deviation is a fairly universal measure of variability

(i) it measures typical variation in the same units and scale as the original data

(ii) it is mathematically convenient, as squares and square roots can effectively be plugged into more complex formulas. Absolute values encounter problems on that front

This is a good moment to attach this link, idea behind pop vs sample var
https://en.wikipedia.org/wiki/Bessel%27s_correction

Variance and Standard Deviation for a Sample

We can estimate the population mean effectively by using the sample mean or the population proportion using the sample proportion.The same is not true for measures of variability.

-The range in a sample (particularly a small one) is almost always going to be biased—it will usually be less than the range for the population Likewise, the sample variance and standard deviation are not the best estimates for the population values because they consistently underestimate the variance and standard devi-ations in the population being sampled.

However, if you divide by n−1 instead of n, the variance and standard deviation from a sample become unbiased estimators of the population values.

***Definition:* Sample variance**

$$\text{Sample variance} = s^2 = \frac{\sum(x-\bar{x})^2}{n-1}$$

**Proof attach :)

Degrees of freedom → number of variables free to vary. *The denominator in the sample variance formula is the number of degrees of freedom

“ Is it really necessary to get into degrees of freedom when first introducing the variance and the standard deviation? I don't think so. It's a strange concept (as Walker, 1940, pointed out many years ago) that students always have trouble with, no matter how you explain it. The number of unconstrained pieces of data? Something you need to know in order to use certain tables in the backs of statistics textbooks? Whatever”

-This is a good article by Professor (?) Knapp :)

http://www.statlit.org/pdf/2013-Knapp-N-vs-N-minus-1-re-visited.pdf

*Attempt :D

Try It Yourself 1.10

In your resampling software, randomly generate a population of 1000 values. It does not matter what population you generate—let us say a population of randomly selected numbers between 0 and 9. In Excel, you can do this with the RANDBETWEEN function. Next, find the variance of this population using the population variance formula. Then, repeatedly take resamples of size 10 and calculate the variance for each resample according to the same population formula. How does the mean of the resample variances compare to the population variance?

Tutorials for this exercise using Resampling Stats for Excel and StatCrunch can be found in the textbook supplements.

For a Box Sampler resampling tutorial based on this exercise, see the file box_sampler_tutorial.pdf.

Distance →

$$\text{Euclidean Distance} = \sqrt{(w1 - x1)^2 + (w2 - x2)^2 + (w3 - x3)^2 + \cdots + (wn - xn)^2}$$

Consider a poll in which respondents are asked to assess their preferences for the musical genres listed below. Ratings are on a scale of 1 (dislike) to 10 (like) and we have poll results from three students (Table 1.3).

TABLE 1.3 Musical Genre Preferences

Person	Rock	Rap	Country	Jazz	New Age
A	7	1	9	1	3
B	4	9	1	3	1
C	9	1	7	2	2

Consider person C. Is she more like person A or person B? Looking at the scores, our guess would be that person C is more like person A. We can measure this distance statistically by subtracting one vector from the another, squaring the differences so they are all positive, summing them so we have a single number, then taking the square root so the original scale is restored.

Test Statistics

A test statistic is the key measurement that we will use to judge the results of the exper-iment or study.

Important: Throughout this example, we will be talking about "reductions in number of errors," not in the number of errors.

Mean reduction in errors (treatment) minus mean reduction in errors (control)

Database Stuff:

Flat File: A flat file is a table that has two dimensions—rows and columns

Relational database: A relational database is composed as a set of tables, each of which has a key column used to relate the information in one table to another.

Database normalization: Normalization of a database is the process of organizing data so that it is stored in a set of related tables with defined linkages.

Structured query language (SQL): SQL is a programming language used to extract information from relational databases and to manipulate the tables in those databases

Big Data: The challenge big data presents is often characterized by the four Vs—volume,velocity,variety, and veracity.*Volume* refers to the

amount of data. *Velocity* refers to the flow rate—the speed at which it is being generated. *Variety* refers to the different types of data being generated (money, dates, numbers, text, etc.).*Veracity* refers to the fact that data are being generated by organic-distributed processes (e.g., millions of people signing up for services or free downloads) and not subject to the controls or quality checks that apply to data collected for a study.

Data Science/Analytics: Both are somewhat new terms and their definitions are hard to pin down. But central to both is the notion of using statistical and machine learning methods to extract useful information from available organizational data (often of huge size).

Variables

Quantitative Variables→ They are an example of a measurement variable or quantitative variable. These are numbers with which you can do meaningful arithmetic. They fall into two types: *discrete* and *continuous*.

> *Def Discrete variable*:The values in a discrete variable differ by fixed amounts and do not assume intermediate values.The most common type of discrete variable is an integer variable, in which only integers are legal values. Family size is an example.
>
> Def:Continuous variable The values in a continuous variable can assume any values and the difference between any two values can be divided up into any number of legal values.Age is a continuous variable as is elevation or longitude or latitude. Often, continuous variables may be binned into discrete variables for convenience.

Categorical Variables→ The other main type of variable is called categorical or qualitative.

> Def:Categorical variable A categorical variable must take one of a set of defined non-numerical values—yes/no,low/medium/high, mammal/bird/reptile, and so on.

Def: Outliers A value (for a given variable) that seems distant from or does not fit in with the other values for that variable is called an outlier. It could be an illegal value, as in this case. It could also be a very odd value or a legitimate one Outliers are not necessarily errors—some are legitimate values. Whenever we find an outlier, we need to *investigate* it and try to understand the reason for it. If there is an error, we need to try to correct it. Outliers, whether erroneous or legitimate,can strongly affect the numbers we compute from our data. In some cases, an outlier is a symptom of a deeper problem that could have an even greater impact on our results.

Histogram- graphical representation of frequency distribution. We group the data into bins. Itis important that the bins be

(i) equally sized and

(ii)contiguous. By contiguous, we mean that the data range is divided up into equally sized bins, even if some bins have no data.

Steam - Life plot: A variant of the histogram, in which the counts of that you have seen earlier are replaced with numbers denoting the actual values

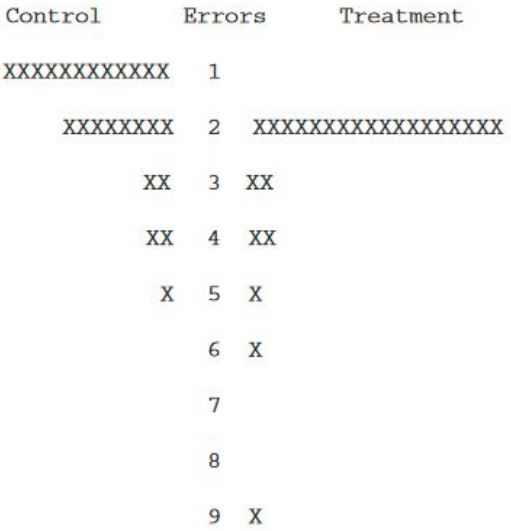


Figure 1.6 Back-to-back histogram.

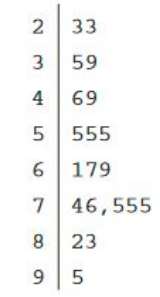


Figure 1.7 Stem-and-leaf plot, hypothetical rural hospital sizes.

Boxplot: Categorical data and its frequency distribution.

- A central box encloses the central half of the data—the top of the box is at the 75thpercentile and the bottom of the box is at the 25th percentile.
- The median is marked with a line.
- "Whiskers" extend out from the box in either direction to enclose the rest of the data or most of it. The whiskers stop when they come to the end of the data or when they get further than 1.5inter-quartile range(IQR), from the top and bottom of the box—whichever comes first.
- Outliers beyond the whiskers are indicated with individual markers

Tail of the distribution:The part of the picture where the data trail off.
Skew: the direction of the longer tail. The shape of the distribution is easier to see in the histogram than in the table.

→ Answer This questions !

- use coin flips to replicate random processes and interpret the results of coin-flipping experiments,
- define and understand probability,
- define, intuitively, *p*-value,
- list the key statistics used in the initial exploration and analysis of data,
- describe the different data formats that you will encounter, including relational database and flat file formats,
- describe the difference between data encountered in traditional statistical research and "big data,"
- explain the use of treatment and control groups in experiments,
- explain the role of randomization in assigning subjects in a study,
- explain the difference between observational studies and experiments.

2

STATISTICAL INFERENCE

Statistical Inference: trying to assess the impact or random variability on the conclusion drawn from a study, or the results of a measurement.

- *Hypothesis Test:* seeks to assess weather the effects we see in some data are real or might just be the result of chance variation.
- *Null hypothesis:* "Could the result be due to chance?" Nothing is happening, and whatever difference is observed is merely the effect of chance.
    - How likely is chance to be the reason in our case (observation)?
    - If our observation is consistent with the range of outcomes that chance might produce, then we associate our results with chance, and fail to reject the null hypothesis.
    - If our observation is an extreme value (a rarity) in the chance model, then we can associate the result not to chance but to a specific factor.

Repeating results: one way to check our result is to repeat the experiment through computer simulation. If our observation, is indeed the fruit of chance (luck of the draw), then multiple simulation of the same experiment should indeed expose this reality: whether the observed is a common value when repeated (say 1000), or whether the observed is not that common in multiple trials.

Permutation test:

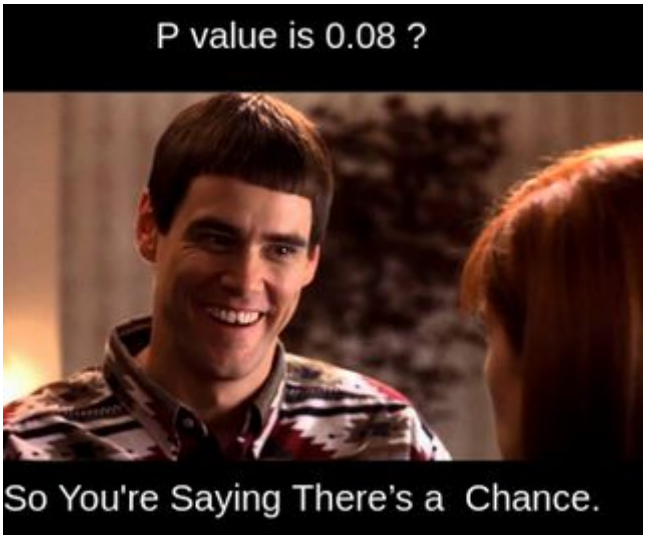
- combining two or more samples
- Shuffling them together
- Picking out resamples at random (or randomization test -exhaust all possible ways)
- Assigning them to a group

P-value: The probability that the chance model might produce an outcome as extreme as the observed value. The more likely under the chance model, the higher the p-value, the less likely to happen under the chance model, the smaller the p-value.

-How rare does rare have to be to reject a chance model? Arbitrary (5%)

- Firm rule for simulations required? Arbitrary (purpose based)

The probability of seeing a result from the null model as extreme as the observed value is called the p-value or probability value.



Exact Tests: examining all of the possible permutations shufflings (exhaust all possible routes).

Alpha: decision threshold you set for the p-value in advance of an experiment (how rare does an event has to be to be able to reject the chance model?)

- Some researchers advocate setting up a level for alpha in advance. The idea is that you set the rules before playing the game. Then, you cannot decide at the end of the game to let alpha be whatever it needs to be for you to win.
- Alpha, should be set by having in mind the trade off between the type I & II errors

Type I Error: "When you erroneously conclude that an effect is real when it is just chance, you have com-mitted a Type I error. This occurs when you get a very low p-value, which indicates a low probability of the result happening by chance, but the result is, nonetheless, still due to chance."

Type II Error: "When you conclude that an effect could be due to chance although it is real, you have committed a Type II error. This occurs when the effect is real, but, due to chance and small sample size, you get ap-value that is not low enough."

Statistical Significance: we are concerned whether a determinate cure is better than another cure. When we are able to reject the null hypothesis, and accept the alternative, we are indeed achieving statistical significance: p-value < alpha ;)

- Statistical significance does not correspond to practical significance, simply finding that the new method is indeed able to yield a difference/betterment, it is

up to experts in that field-domain to determine whether such difference is worth pursuing.

Effect size: how big the difference is between H0 and Ha in terms of the context of the problem

Exercise Simulation + Review Questions :)

- explain the concept of a null hypothesis,
- describe how to conduct a permutation test with a hat and slips of paper,
- interpret the results of a permutation test,
- describe the shape of the Normal distribution and what is meant when it is said that a more accurate name is the "Error" distribution,
- define, in the context of hypothesis testing, alpha, Type I error, and Type II error,
- explain in what circumstances hypothesis testing is used.

3

DISPLAYING AND EXPLORING DATA

Pie Charts: are not that good in terms of presenting information in a clear way, given that human perception is better suited at comparing lengths, instead of angles.
Misuse of graphs: time periods can have an effect on the inference of the data.
**This chapter did not have much useful information,* hence I will not do the review q's

4

PROBABILITY

Probability: two useful interpretations

- Long-run frequency, if we repeated an experiment over and over, the expected mean value
- Degree of belief, lack of a repeatable process, but idea of probability still relevant

Sample space: the list of all possible outcomes of a specific event

- We say the outcomes must be jointly exhaustive and mutually exclusive
- Jointly exhaustive & mutually exclusive

Operations:

- Complement
- Intersection
- Union
- Disjoint

Random variable: variable (attribute) that takes on different values as a result of a random process

Weighted mean: each value comes with a weight (percentage/proportion)

- Multiply outcome * weight

Expected value: used to evaluate a possible future outcomes and their probabilities when there is no prior set of data to take the mean, where data comes from the 'best' data the analyst has.

Standardization (Normalization): standardization technique that aims at putting all of the data on the same scale (x-mu)/sigma. This process, takes units away from the data, and allows us to make comments regarding an observations in terms of SD and mean.

Standard Normal Distribution:

- mean 0, ,sd=1.
- The z-scores are on the x axis
- the total area under the curve is equal to 1
- 68%, 95%, 99.7%

5

RELATIONSHIP BETWEEN TWO CATEGORICAL VARIABLES

Simpson's Paradox: how can a group have high rates in different categories, but fail to have high overall rates? 'Simpson's Paradox is one of a family of paradoxes or oddities that results from aggregation—putting parts together into a whole. In general, putting parts together results in a whole that looks like the parts, all other things being equal. Simpson's Paradox arises when all other things are not equal.'

Comparing Proportions: proportion (parts/whole) of one group against another and perform an hypothesis test (Ho, Ha).
Check out example on github
https://github.com/dasvidanja/Statistics/blob/master/Statistics%20Fun/Proportion_Difference_Hypothesis_Testing.ipynb

Conditional probability:

$$P(A|B) = \frac{P(A \cap B)}{P(B)}$$

Baye's Rule

$$P(A|B) = \frac{P(A)\,x\,P(B|A)}{P(A)\,x\,P(B|A) + P(\sim B|\sim A)}$$

General Independence:

$$P(A \cap B) = P(A) * P(B|A)$$

Independence

$$P(A \cap B) = P(A) * P(B)$$

6

SURVEYS AND SAMPLING

Bias: whenever we apply a statistical procedure or measure to a sample from a population, and it *consistently* produces overestimates or under-estimates of a characteristic of that population.

From Gallup's Story: small representative sample is more accurate than a large sample that is not representative.

Simple Random Sampling (**SRS**)**:**

- Produced by first placing the entire population (slips of papers) in a box
- Shuffle box, and draw out enough observations for a sample
    - Key idea: a sample of size n qualifies as SRS if each combination of n elements there is an equal chase of emerging as the selected sample.
    - Focus: procedure by which sample is drawn (not characteristics of resulting sample
- Random sampling does nor guarantee a completely representative sample (guarantees that each sample will be a little different from its population).

Population: group that we are studying (never really known)
Sampling Frame: slips of paper in a box, database, list of names, …
Parameter: measurable characteristic of a population
Sample: subset of population- random sample if randomly selected
Random Sampling: a process in which each element of a population has an equal chance of being selected.
Statistics: measurable characteristics of a *sample*

Random Assignment—Review

We have been talking thus far about the use of random sampling from a larger population to form representative samples. In the first course of this sequence, we spoke about the use of random *assignment* of treatments to subjects in experiments. The mechanics are similar—in each case, we can imagine a box with slips of paper and a random draw procedure.

In the case of the political survey, the goals of random sampling are to produce a sample that is reasonably representative of a larger population and to estimate the extent to which a sample estimate might be in error due to chance.

In the case of the experiment, the goal of random assignment is to ensure that any difference between the treatment groups is either due to the treatment being tested or to chance.

Margin of Error: quantifies the extent to which a sample might misrepresent the population it is coming from.

Confidence Interval: we can quantify the uncertainty of the range of the resamples results with an interval that includes the large majority-90 95% ….

Sampling with replacement is **equal** to sampling without replacement from a huge population.

Summing Up

To produce a confidence interval,

1. We can use the observed sample as a good proxy for the population.
2. The resample size should be the same as the original sample size.
3. The fact that the sampling is done *with replacement* allows the sample to serve, in effect, as a simulated population of infinitely large size.

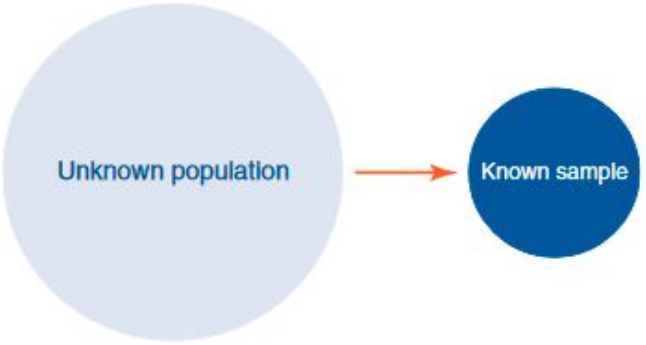
Bootstrap: sampling with replacement from a hypothetical population (available sample) of size n; get a statistics for all resamples, get the average of that statistics from all samples. Goal: get distribution of a determinate statistic.

Observation: single case in our data
Sample: collection of observations from a population
Resample: new simulated sample drawn/or generated randomly from original sample
Sampling with Replacement: item is replaced (with itself) after it is drawn
Sampling without Replacement: a.k.a shuffling, once an item is drawn it cannot be drawn again

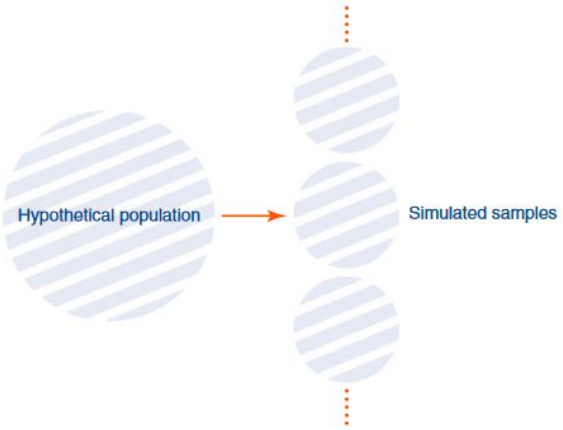
Single simulation trial: taking a resample and calculating a statistic (mean, median..)
Simulation: repeat of single single simulation trials & collection of their calculations

Let's Recap

We have a sample of size *N* from an unknown population and we want to know how much an estimate based on that sample might be in error.



The key question is how do samples drawn from this population behave, that is, how different are they from one another? We address this by simulating resamples of size *N* from a hypothetical population.



Bootstrap STEPS:

1. From our observed sample, calculate a statistic from it to measure some attribute of the population that we are examining.
2. Create a hypothetical population, incorporating the best information we have. This is usually the information from the sample.
3. Draw a resample from the hypothetical population and record the statistic of interest.
4. Repeat step 3 many times.

*At the end observer the sampling distribution of the statistic to check out how much of our original estimated (step 1) might be in error /vary

Types of Sampling:

Stratified Sampling: population is split into categories, or strata and separate samples are drawn from each stratum. Ensure that we have equal sized samples from each stratum.

Cluster Sampling: clusters of subjects are selected, and the subjects within those clusters are surveyed. This is practical and efficient

Systematic sampling: the selection of every nth record. Pay attention to bias (say daily sales, if you measure sales only at a certain hour or day, you may not be representing the truth)

Multistage sampling: ideally you want to minimize cost, sampling error, and bias.

Convenience sampling: no effort to define a population, nor no effort to ensure sample is representative of the population. Easy, cheap, but not consistent.

Self- selection: this guarantees bias-given that only individuals with strong opinions will participate in the survey.

Non response Bias: can occur in any sampling method. The idea is that the opinion of those who respond to the survey is different than the opinion of those who do not respond to the survey.

7

CONFIDENCE INTERVALS

- distinguish between the appropriate uses of point estimates and interval estimates,
- calculate confidence intervals (via resampling or formulas),
- explain the relationship between the Central Limit Theorem and the applicability of Normal approximations for confidence intervals,
- calculate standard error and explain the difference between it and standard deviation,
- calculate the confidence interval for a mean or proportion,
- calculate the confidence interval for a difference in means or proportions.

This material, particularly the vocabulary and definitions, is most relevant for the *research* community. *Data scientists*, however, will encounter confidence intervals in their work and will benefit from a solid understanding, via resampling, of how they work.

**Code Problem with bootstrap*

Resampling Procedure (Bootstrap):

1. Write all 20 sample values on slips of paper and place them in a box.
2. Draw a slip from the box, record its value, and replace the slip.
3. Repeat step two 19 more times and record the mean of the 20 values as shown in Figure 7.1.
4. Repeat steps two and three 1000 more times.
5. Arrange the 1000 resampled means in descending order and identify the fifth percentile and the 95th percentile—the values that enclose 90% of the resampled means. These are the endpoints of a 90% confidence interval, as shown in Figure 7.2. Figure 7.3 is a histogram of the 1000 resampled means.

Point Estimates: the process of establishing the possible error is a confidence error: one way to measure the accuracy of a measurement. The statistic itself is called a point estimate. A point estimate is a single value (not a range), an estimator of some parameter for a population from a sample of that population.

Confidence interval: given that whenever we are calculating estimates from one sample to another, the resulting estimate tends to fluctuate from one point to another. As a result, it is more appropriate to provide a range of the possible estimates. The uncertainty is quantified with a confidence interval.

https://www.datasciencecentral.com/profiles/blogs/significance-level-vs-confidence-level-vs-confidence-interval

Significant level: alpha, is probability of rejecting ho when ho is true (making the wrong decision when the null is true)
Confidence level: this regards the procedure, the probability of getting the same results if we repeated the procedure over and over
Confidence interval: range of values expected to contain the population parameter of interest.
"A more technical definition of a 90% confidence interval is that it is an interval that would enclose the true statistic 90% of the time when constructed repeatedly in the same manner with the same population."

CI vs Margin of Error:

-Margin of error is simply +- quantity attached to the estimate
-CI, is the actual endpoints of the interval
- First we compute CI, and then we calculate the margin of error (if symmetric)

7.3 CONFIDENCE INTERVAL FOR A MEAN

We have actually already calculated a confidence interval for a mean, in effect. Let us review the Toyota Corolla case described earlier, where the average price in the sample—the sample mean—is €17,685.

Resampling Procedure (Bootstrap):

1. Write all 20 sample values on slips of paper and place them in a box.
2. Draw a slip from the box, record its value, and replace the slip.
3. Repeat step two 19 more times and record the mean of the 20 values as shown in Figure 7.1.
4. Repeat steps two and three 1000 more times.
5. Arrange the 1000 resampled means in descending order and identify the fifth percentile and the 95th percentile—the values that enclose 90% of the resampled means. These are the endpoints of a 90% confidence interval, as shown in Figure 7.2. Figure 7.3 is a histogram of the 1000 resampled means.

Formula-Based

Back in the day computers were not as prominent as today, as a result, instead of simulations, statisticians developed approximations that allowed them to calculate confidence intervals from formulas.

- Binomial
- Normal distribution

Central Limit Theorem

Despite the original population not having a normal distribution, the normal distribution may still apply for the means of samples drawn from the population.

- Depending on the size of the sample (20-30)
- The degree of non-normality in the parent population

Definition:

"The Central Limit Theorem says that the means drawn from multiple samples will be Nor-mally distributed, even if the source population is not Normally distributed, provided that the sample size is large enough and the departure from Normality is not too great."

Assumptions:

1. Randomization condition (data must be sampled randomly)
2. Independence (the occurrence of one event has no effect on the occurrence of another event
3. 10% condition (sample size should be no more than 10% of population, if drawn without replacement
4. Sample size should be sufficiently large (if population is skewed than we need a larger sample than we would, if the population was normally distributed , n=30 usually)

Z intervals for a Mean

1. Find values in the Standard Normal distribution that correspond to z-interval (90,95…)
2. Multiply by sample standard deviation (s) (divide) square root of sample size (n)

$$\overline{x} \pm z\frac{s}{\sqrt{n}}$$

T-interval for a mean: whenever our sample size is smaller than 30, the t distribution is a better approximation. The t-distribution is the same as the normal for sample size> 30. As sample size diminishes, the t- distribution takes into account the volatility that comes with smaller sample sizes.

1. Specify degrees of freedom (n-1)

$$\text{The } 100*(1-\alpha)\%\ t\text{-interval for the mean} = \left(\bar{x} - t_{n-1,\,\alpha/2} * \frac{s}{\sqrt{n}}, \bar{x} + t_{n-1,\,\alpha/2} * \frac{s}{\sqrt{n}}\right)$$

Standard Error

Similar to standard deviation. Both measures spread. Standard error uses statistics, standard deviation uses parameters. The standard error tells you how far the sample statistics deviates from the actual mean. SE is used to measure the statistical accuracy of an estimate. Bottom line se gauges accuracy of an estimate, whereas sd is the measure which assesses the amount of variation within a set of observations.

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

BASIS FOR COMPARISON	STANDARD DEVIATION	STANDARD ERROR
Meaning	Standard Deviation implies a measure of dispersion of the set of values from their mean.	Standard Error connotes the measure of statistical exactness of an estimate.
Statistic	Descriptive	Inferential
Measures	How much observations vary from each other.	How precise the sample mean to the true population mean.
Distribution	Distribution of observation concerning normal curve.	Distribution of an estimate concerning normal curve.
Formula	Square root of variance	Standard deviation divided by square root of sample size.
Increase in sample size	Gives a more specific measure of standard	Decreases standard error.

Confidence intervals for a single proportion

Resampling Steps

We want to determine how samples of size 20 might differ from one another. We do not know what the returned proportion is in the population, so we use the sample proportion, 0.20, as our best guess. Then, we draw resamples from a population box that contains 20% returned. The products being sold by the company, worldwide and over time, constitute a very large population, much larger than 20, so we resample with replacement. This reflects the fact that the probability that a product will be returned is constant from one sale to the next.

(1) We can represent the 20% returned by a box with four ones, that is, returns, and 16 zeros, that is, not returned.
(2) Draw a number from the box and record whether it is a zero or a one. Replace the number.
(3) Repeat step two 19 more times, for a resample of 20 from the box. Record the number of returns—ones.
(4) Repeat steps two and three 1000 times, recording the number of ones in the resample each time.
(5) Order the results and find the fifth and 95th percentiles. This is a 90% confidence interval for the number of products returned. Divide by 20 and multiply by 100 to get percent returned.

Binomial Distribution:

For small samples we can use the BD to determine probabilities of determinate discrete results. For larger results, we can use its normal approximation.

1. Assumption: independence

Normal Approximation:

In cases where we have large samples sizes, than the binomial calculations become too time I(computationally) consuming. Hence, we use the normal distribution approximation which yields similar results.

Confidence interval for a difference in means

"The confidence interval procedure asks "How would this result differ if we drew many additional samples?"

Resampling Procedure—Bootstrap Percentile Interval

In constructing a confidence interval for this problem, that is, to have the resampling world replicate the real world, we will want to resample from ***two*** boxes—one box for vendor A and one box for vendor B. (In the next chapter, we will see that if we were doing a hypothesis test, we would want just one box representing the imaginary world of a single population whose behavior we would want to test.) We will calculate a 90% confidence interval.

1. Box A has 12 slips of paper with the 12 values for Vendor A
2. Box B has 10 slips of paper with the 10 values for Vendor B
3. Draw a sample of 12 with replacement from Box A and record the mean.
4. Draw a sample of 10 with replacement from Box B and record the mean.
5. Record the difference—Mean A minus Mean B
6. Repeat steps three through five 1000 times.
7. Review the distribution of the 1000 resampled means by creating a histogram and find the 5th and 95th percentiles. These are the bounds of a 90% confidence interval. See Figure 7.6 for the histogram our example produced. Specific software procedures for this example using Resampling Stats, StatCrunch, and Box Sampler can be found in the textbook supplements.

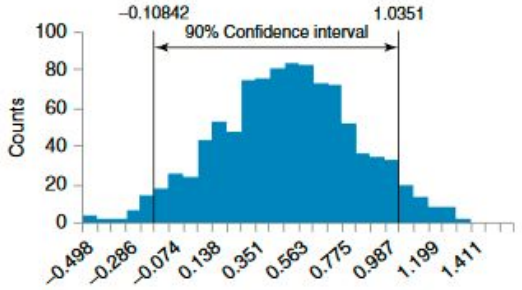


Figure 7.6 Histogram with 90% confidence interval—mean of A minus mean of B.

Formula—Confidence Interval for a Difference in Means

The conventional procedure for drawing a confidence interval around the difference in means uses the same t-distribution developed by Gossett that we saw.

We will denote the sample sizes of the two samples as n_1 and n_2, the sample means as $\overline{X}_1$ and $\overline{X}_2$, and the sample variances as s_1^2 and s_2^2. We further assume that the two samples are independent—not paired. The lower and upper bounds of a $1-\alpha$ confidence interval are calculated below.

Lower bound:

$$\overline{X}_1 - \overline{X}_2 - t_{v,\,\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Upper bound:

$$\overline{X}_1 - \overline{X}_2 + t_{v,\,\alpha/2} * \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

The degrees of freedom of Student's t-distribution are represented by v, the Greek letter nu. Degrees of freedom can be found with this equation.

$$v = \min(n_1, n_2) - 1$$

You may also see references to calculations using a "pooled variance." Such calculations assume that the two samples, ostensibly coming from two different populations, share the same variance. The circumstances under which this would be true are sufficiently limited that the pooled variance case is not covered here.

Confidence intervals for a difference in PROPORTIONS

Resampling Procedure

Because Box Sampler is limited to a sample size of 200, this simulation must be done using either Resampling Stats for Excel or StatCrunch.

1. In one box—the high cholesterol box—put 10 slips of paper marked 1 for heart attacks and 125 slips marked 0 for no heart attacks.
2. In a second box—the low cholesterol box—put 21 slips of paper marked 1 and 449 slips marked 0.
3. From the first sample, draw a resample of size 135 randomly and with replacement. Record the proportion of ones.
4. From the second sample, draw a resample of size 470 randomly and with replacement. Record the proportion of ones.
5. Record the [result from step three] minus the [result from step four].
6. Repeat steps three through five 1000 times.

7. Find the interval that encloses the central 95% of the results—chopping 2.5% off each end. Figure 7.7 illustrates this interval. Specific software procedures for this example using Resampling Stats and StatCrunch can be found in the textbook supplements.

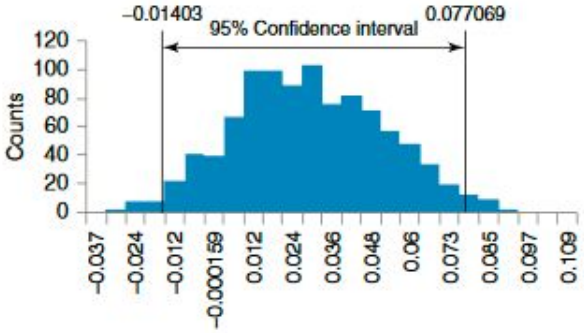


Figure 7.7 Histogram with 95% confidence interval, difference in proportion 1s, resample group of 135 minus resample group of 470.

7.9 RECAPPING

The vocabulary and details of statistical inference can be confusing and appear disconnected from applications, so this chapter merits a brief summary:

1. Learned three roughly equivalent ways to calculate a confidence interval for a single proportion:
   - Resample 0s and 1s from a box, recording the proportion of 1s each time.
   - Use the binomial formula.
   - Use the Normal approximation to the binomial.

   All three methods are most likely to be deployed via software; details of the binomial formula and the Normal approximation have been described in an appendix.
2. Learned two roughly equivalent ways to calculate a confidence interval for a difference between two means:
   - Resample the actual values separately from two boxes, recording the difference in resample means.
   - Use the *t*-distribution.
3. Learned how to calculate a confidence interval for a difference in proportions via resampling:
   - Resample 0s and 1s from two separate boxes, recording the difference in proportion of 1s.

8

HYPOTHESIS TESTS

BASIC TERMINOLOGY

Statistical inference: process of accounting for random variation in data as you attempt to draw conclusions from it.

CI & HI role: tools to not get fooled by chance!

Confidence interval: 'How much change error might there be in this measurement or estimate or model, owing to the luck of the draw in who/what gets selected in a sample?'

Hypothesis test: 'Might this apparently interesting result have happened by chance, owing tot he luck of the draw in who/what get selected in sample or assigned to different treatments.' We have this imaginary null model, we draw resamples to see whether we could get outcomes as extreme as the observed outcome.

Null model: imaginary chance model representing the idea that nothing new or novel is going on or that there is no difference between treatments A and B

EXAMPLE- to code and post on GIT

P-value: the frequency with which a result as extreme as the observed result occurs just by chance, drawing from the null hypothesis model.

Significance or Alpha level: how unusual is too unusual to be ascribed as chance. If the result is smaller than alpha (i.e p-value < alpha) then the results are deemed statistically significant. (In the DS community result has indeed information value, but it is not of critical decision importance).

Test Statistic: A test statistic is the key measurement that we will use to judge the results of the exper-iment or study.

> **Critical value:** what value of the test statistic corresponds to a given alpha level (cut off for chance vs not chance)

A-B Test - two sample comparison

The two-sample comparison is a fundamental inference procedure in statistical analysis. It basically asks the question: "Is treatment A different from treatment B?"

Basic Two-Sample Hypothesis Test Concept

1. Establish a null model, which is also called *the null hypothesis*. This represents a world in which nothing unusual is happening except by chance. Usually, this null model is that the two samples come from the same population.
2. Examine pairs of resamples drawn repeatedly from the null model to see how much they differ from one another. Alternatively, we can use formulas to learn about this distribution of sample differences. If the observed difference is rarely encountered in this chance model, we are prepared to say that chance is not responsible.

The procedure

Resampling (point 1 important)

Basic Two-Sample Hypothesis Test Details

1. Make sure you clearly understand
   - the sizes of the two original samples,
   - the statistic used to measure the difference between sample A and sample B, for example, the difference in means, proportions, and ratio of proportions
   - the value of that statistic for the original two samples.
2. Create an imaginary box that represents the null model. Examples could be a box with eight red chips and two black chips (to represent something with a 20% probability) or a box with all the observed body weights (in a study of how something affects body weight), where each weight is written on a slip of paper.
3. Draw out two resamples of the same size as the original samples. This can be done with or without replacement. The two procedures yield similar results, but do diverge for very small samples (<10), and both are used. The distinction between them is technical and beyond the scope of this course.
4. Record the value of the statistic of interest.
5. Repeat steps 3 and 4 many times for 1000 trials. Even more trials can be conducted for greater accuracy.
6. Note the proportion of trials that yields a value for the statistic as large as that observed.

Formula approach

Formula approaches are not as flexible as resampling techniques, and more complex.

1. Comparing two means
    a. T test (sample size at least 15)
        i. Pooled, assume the samples come from the same population (var1 == var2)

$$t = \frac{\overline{X}_1 - \overline{X}_2}{S\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

$$S^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}.$$

        ii. Unpooled
    b. Z test (sample size at least 30)
        i. Pooled
        ii. Unpooled

2. Comparing two proportions

3. Paired Comparisons

Example for the PIGs

TABLE 8.1 Blood Loss in Pigs (ml)

Control Group	Treatment Group
786	543
375	666
4446	455
2886	823
478	1716
587	797
434	2828
4764	1251
3281	702
3837	1078
Mean: 2187	Mean: 1086

Difference in the means of treatment minus control = −1101.

Resampling Procedure

1. Write blood loss from each pig on a slip of paper and put all 20 slips of paper in a box.
2. Shuffle the box and randomly draw with or without replacement two resamples of 10 each.
3. Find the mean of each resample, subtract the mean of the first resample from the mean of the second, and record this difference.
4. Repeat steps two and three 1000 times and find out how often the recorded difference is ≤ -1101 ml.

Null & Alternative

In hypothesis testing, the goal is to determine whether chance is responsible for an effect, or whether that effect is indeed real. One key part of our process is the NULL model

Paired Comparisons

In our standard two-sample comparison, the differ-ences get swamped by the variation among subjects. The standard test, whether a resampling or at-test, cannot discern the treatment effect in the face of all the between-subject variation. In statistical terms, it lacks power. Null model that does not assume two different groups belong to the same population.

Power: Power is the probability that a statistical test will identify an effect, that is, determine that there is a statistically significant difference when one exists.

To calculate power, you need to know

(i) the effect size you want to discern,

(ii) the sample sizes,

(iii) something about sample variances and distribution.

Z vs T test

Z test	T test
- Know population variance - Sample size >=30 * Sample variance can be a good approximation for the population variance	- Unknown population variance - Sample size < 30 * T-test is the same for n>30

from :

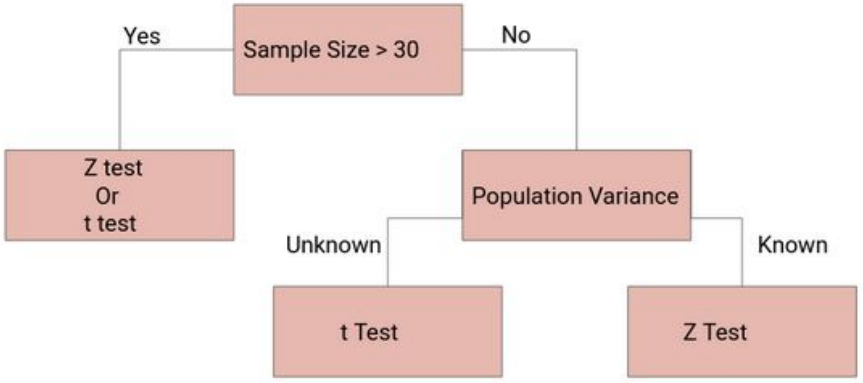
https://www.analyticsvidhya.com/blog/2020/06/statistics-analytics-hypothesis-testing-z-test-t-test/

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sigma\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

(if population variance is known)

Deciding between Z Test and T-Test

So when we should perform the Z test and when we should perform t-Test? It's a key question we need to answer if we want to master statistics.



One-Way or Two-Way Hypothesis Test

The Rule

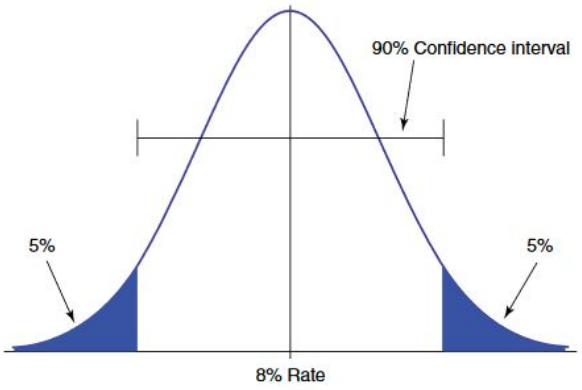
"If the alternative hypothesis is unidirectional, perform a one-way or a one-tailed test. In other words, if the original study requires a difference in a specific direction, count only those differences "

An example of a unidirectional—one-way—hypothesis is the question of whether a treatment is better than the control

"If the alternative hypothesis is bi-directional, then the differences in either direction should be counted. Such a two-way test is exemplified by the question of whether adver-tisement A or B is better."

CI vs Hypothesis Test

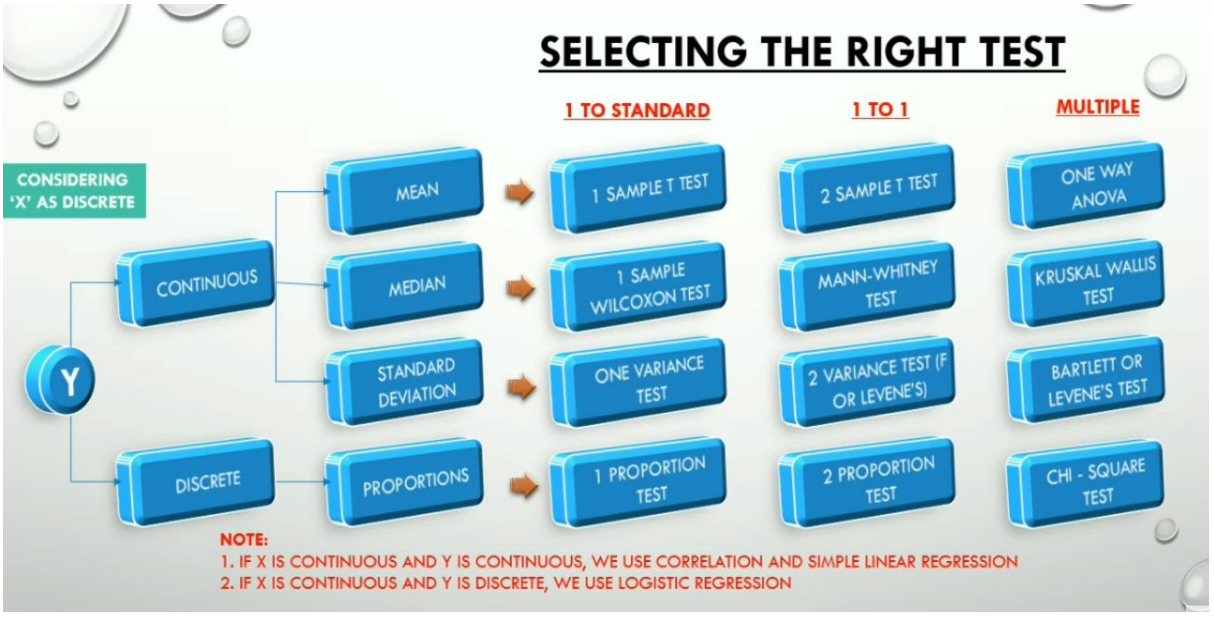
CI: how inaccurate a measurement/statistic might be, based solely on the luck of the draw in what gets selected. No null model, just a hypothetical population which we draw samples from and compare the likelihood of the computed statistics over a large number of samples (does it vary a lot, or is it stable).



9

HYPOTHESIS TESTING—2

Perhaps, an overview from the websites 6sigma.com



Whenever we have more than 2 groups:

- Discrete data → Chi Square
- Continuous data → ANOVA

Goodness of fit

"It examines how well an observed distribution fits a theoretical expectation."

Code example ?