MATH-338 Midterm 2 Cheat Sheet

THEORY

Day 14: probability density function is represented an integral with function $f(x)$. Our probability lies within the curve and is always 1. Density curve $\rightarrow$ bell curve. Z-Score allows us to have a universal standard for density curves with different scales. They are directly proportional to the standard deviation and the delta from the mean of the graph.

Day 15: unimodal: one hump, bimodal: two humps. Mean is resistant whereas the mean is subject to change. Density curves decay to histograms (integral $\rightarrow$ to Reimann Sum). Whisker plots are an effective method to determine if a data set contains outliers (data points not belonging to the sample set)

FORMULAS

- $\sqcap = width \times \frac{1}{width}$ (finite curve)
- $Z = \frac{x-\mu}{\sigma}$ (z-score)
- $X \sim N(\mu\,, \sigma)$

- IQR $= Q_3 - Q_1$
- $K = 1.5$
- Lower fence: $Q_1 - K \times IQR$
- Upper fence: $Q_3 + K \times IQR$

Day 1

Statistics

A set of tools for understanding data and making decisions/conclusions/predictions under uncertainty

Randomness

- in the short term, we don't know what will happen (flipping a coin)
- in the long term, we know the **distribution** of possibilities (what outcomes are possible and how often they occur)
  - the crux of randomness
  - as the amount of times we get a random variable approaches infinity, the clearer/less random the variable becomes

Two definitions of probability

- **proportion** of times an outcome occurs or would occur over infinitely many repetitions of a random action
  - Frequentist
  - math is a lot nicer
  - there is a fixed outcome but we don't know it
- a number quantifying our **belief** that an outcome can/will occur
  - Bayesian
  - 2000 times more intuitive and math is just as hard
  - random outcome (not fixed)

Both are calculus based

Probability Model

Consists of two parts:

- sample space: list (set/list of all unique values) of all possibilities and must be well defined.
- probability of each outcome

This in essence is a hash table

An **event** is an arbitrary set of 0 or more outcomes in a sample space

Axioms of Probability

- axioms : something is so obvious it does not need to be proven
- For events A and B in the same sample space denoted as "S":
  - The probability of event A, denoted as P(A) is a number between 0 and 1 (inclusive).
    - [0, 1] notation as well.
    - **NOTE**: P(A) = 0 means A is "impossible" and P(A) = 1 means A is guaranteed
  - P(S) = 1
    - **Some outcome is bound to happen**
  - If A and B are disjoint (there are no common outcomes. A is not in B AND B is not in A), then P(A or B) = P(A) + P(B)

Simple rules that follows from the Axioms

- Compliment Rule: Define A^C = A compliment, that is A^C is the event "A does not occur"
  - $P(A^C) = 1$ - P(A)
  - The summation of all events that **do not** occur minus the overall probability (100%)
- General addition rule: Suppose events A and B have at least one common outcome
  - Define $A \cap B$ to be the set of outcomes common to A & B
  - Define $A \cup B$ to be the set of outcomes in A, or in B or in both A & B
    - Then $(P(A \cup B)) = P(A) + P(B) - P(A \cap B)$
    - This is the same as this:

```
a = [1, 2, 3]
b = [2, 4, 5]
c = a + b
# c = [1, 2, 2, 3, 4, 5]
c = set(a+b)
# c = {1, 2, 3, 4, 5}
```

Example

Random phenomenon: Draw 1 tile from a standard Scrabble bag of 100 tiles

- Sample space 1 (option one):
  - S = the 100 tiles in the bag
  - All tiles are equally likely to be drawn
  - P(draw particular tile) = 1/100 or 0.01 for all
- Sample space 2 (option two):
  - The 27 "letters" (26 letters and 1 blank)
- Let event C = "draw a letter in CAT"
- Let event D = "draw a letter in PET"

P(C) = P(C) + P(A) + P(T) = .02+.09+.06 = 0.17

P(D) = P(P) + P(E) + P(T) = 0.2+0.12+0.06 = 0.2

$P(C^C)$ = P(do not draw any of the letters in CAT) = 1 - P(C) = 1 - .17 = .83 $P(C \cap D)$ =>

- = P(letter in both CAT & PET)
- = P(T) = 0.06

$P(C \cup D)$ = P(letter in CAT or PET or both words) =>

- = P(C) + P(D) - $P(C \cap D)$
- = 0.17 + 0.2 - 0.06 = 0.31

Python Code Representation

```
#!/usr/bin/env python3.5

# probability can be calculated by using a hash table in conjunction with a set
# hash tables are used when there are two different letters with the same probability
# using a bare list would result in incorrect calculations of probability
# they would be treated as non unique instances
# in turn allowing for it to filter out needed objects
# this boils down to a set of unique hash tables and summing up their values

class hashabledict(dict):
    def __hash__(self):
        return hash(tuple(sorted(self.items())))
value_mapping = {
    "c": 0.02,
    "a": 0.09,
    "t": 0.06,
    "p": 0.02,
    "e": 0.12
}

def get_probability(*args):
    # s has extra new line for code to fit
    s = set((hashabledict({letter: value_mapping[letter]})
        for argument in args for letter in argument))
    return sum([sum(dictionary.values()) for dictionary in s])


print(get_probability("cat", "pet"))
# this some times yields 0.310000000000005 and 0.3099999999999994
# which is essentially the same number
```

vim regex

```
# this replaces all caps in proper latex
:%s/cap/\$\\cap\$/g
```

Outline

1. Recap of probability
2. Simulation
3. Random Variables

Independent vs. Disjoint Events

- Independent events **can** happen at the same time, but knowing that event "A" occurred **does not** change P(B) and vice versa
- Disjoint events cannot happen at the same time.
  - Knowing that event "A" occurred changed P(B) = 0 and vice versa
- If A and B are independent, $P(A \cap B) = P(A)\ P(B)$
- If A and B are disjoint, $P(A \cap B) = 0$

Example One

- Draw a tile from a bag of 100 scrabble tiles
- Event C = "the tile is a C"
- Event A = "the tile is an A"

P(C) = .02

P(A) = .09

Events "C" and "A" are disjoint

Events "C" and "A" are <u>not</u> independent

Example Two

Draw one tile and set it outside

Event C = "first tile is a C"

Event A = "second tile is a A"

Event C and A are not disjoint

Events "C" and "A" are <u>not</u> independent

Sampling without replacement

Example Three

Draw a tile, put it back in the bag and then draw another tile

Event C = "first tile is a C" Event A = "second tile is an A"

Event C and A are <u>not</u> disjoint Event C and A <u>are</u> independent

Sampling with replacement

Simulation

Trying to imitate in the real world where the outcome is uncertain but is random

- Specify our model for an uncertain situation/random event
- "Randomly" generate an outcome for the model
- Repeat step two many, many times

Why simulate?

- Once we set up the model, the math maybe too difficult
- Situation may be unique, or we only have ability to observe it once, due to physical/financial limitations
- For fun and/or profit

Report assumptions of the model!

Random Variables (RVs)

Random variable is a variable whose numerical values describe outcomes of a random event

Typically we map outcomes in our sample space denoted as "S" to numerical values of the random variable.

Discrete Random Variable : probability mass function (PMF) places positive probability at specific numbers on the number line

- Only specific numbers
- Example: all outcomes are real, positive numbers

Continuous Random Variable : probability density function (PDF)

- Places positive probability along a possibly infinite interval of the number line.

Writing the PMF of a Discrete Random Variable

Each unique key value X=x is mapped to an non unique value P(X=x)

```
example_hash_map = {
    key: value
}
```

A hash table is another way to represent data mapping.

- value represents a random variable
- key represents a “realization” of value

Example One

We can find P(Y=0)

Once we have observed the random event either y = 0 or y != 0

Let X = the point value of the chosen tile

```
map = {
    0: 0.02,
    1: 0.68,
    2: 0.07,
    3: 0.08,
    4: 0.10,
    5: 0.01,
    8: 0.02,
    10: 0.02
}
```

Sum of values in map == 0

Example Two

Use PMF & probability rules to find:

- P(X <= 3)
  - P(X = 0, 1, 2 or 3)
  - P(X = 0) + P(X = 1) + P(X = 2) + P(X = 3)
  - = 0.85
- P(X > 1)
  - P(X = 2, 3, 4, 5, 8, or 10)
  - 1 - P(X <= 1) = 1 - (X = 0 or 1)
  - 1 - [P(X = 0) + P(X = 1)] = 1 - [0.02 + 0.68] = 1 - .7 = 0.3
- P(X > 5)
  - P(X = {0..5})
  - 0.04
- P(3 < x <= 5)
  - 0.11
  - P(X = 4, or 5)

Expected Value (Mean) of a Random Variable

- Called expectation, mean, all the same thing
- On average, what value do we expect the random variable to be

Recall idea of "weighted average"

Mean of a Probability Distribution

μ

Denotes the average of all events, which in turn gives us an expected value for a given function.

Summation notation

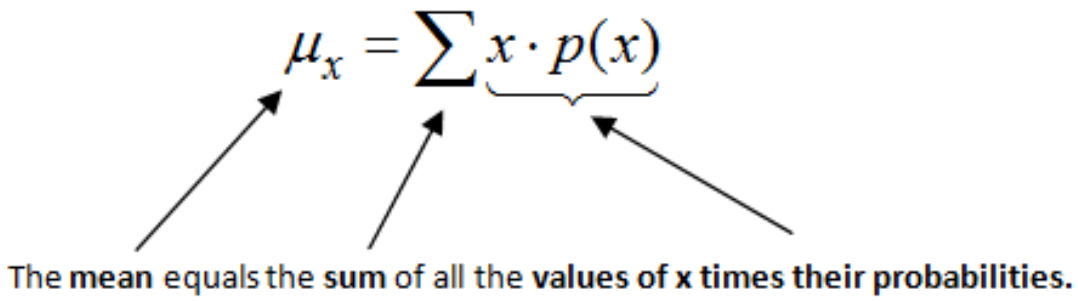


Figure 1: discrete random variable formula

Expected value is a linear operator (can take in sum and give back a result in the form of a sum of the applied operators)

For random variables X and Y, and constant C

- E[X+Y] = E[X] + E[Y]
- E[cX] = cE[X]
- ^ where "c" is a constant applied

This implies, for X, Y and arbitrary constants a,b E[aX + bY] = aE[X] + bE[Y]

Consequences: a = 1 , b = -1

E[X - Y] = E[X] - E[Y]

$\mu_{x-y} = \mu_x - \mu_y$

Day 3

Outline

1. Expected value of the random variable
2. Variance and standard deviation of random variable
3. [If time allows]

Expected values (mean)

Please refer to day_two.pdf

Law of Large Numbers

Suppose we have "N" independent and identically distributed (IID) realization of "X".

That is, we observe out random event "N" times independently and record the value of "X".

Then, as "N" increases, the sample mean of the "N" independent observations converges to μ_x

We can get arbitrarily close to μby simply observing values of "X" enough times.

Variance of a Random Variable

Average squared deviance from mean (distance away from the middle)

Variance Formula:

$$\sigma^2{}_x = \sum[x^2 * P(x)] - \mu_x^2$$

Figure 1: variance formula

- Variance is non-negative
- Variance is not a linear operator

In general, Var(X+Y) != Var(X) + Var(Y)

However, if X and Y are independent

Var(X+Y) == Var(X) + Var(Y)

Var(cX) != cVar(X)

However! –> Var(cX) == C^2Var(X)

When X and Y are independent,

Var(aX+bY) = $a^2\sigma^2{}_x + b^2\sigma^2{}_y$

Standard Deviation of Random Variable

$\sigma_x = \sqrt{\sigma_x}$

Standard deviation is not linear

$\sigma_{x+y} = \sqrt{\sigma_x + \sigma_y}$

If X and Y are independent

$\sigma_{cx} = |C|\sigma_x$

Adding a constant

Consider W = X + c (where c is an arbitrary constant)

$\Sigma[W] = \Sigma[X+c] = \Sigma[X] + \Sigma[c]$

$\Sigma(W) = \Sigma(X) + c$

$Var(W) = Var(X+c)$

$= Var(X) + Var(c)$

$Var(c) = 0$

$Var(W) = Var(x)$

$SD(W) = SD(X)$

Example One

- Toss two fair coins.
- Let X be the number of heads observed
- Find the PMF, expected value, variance and standard deviation of X.

Each win is independent

P(Heads) = 1/2

Independence: $P(A \cap B) = P(A) * P(B)$ 0 heads: TT => $P(TT) = P(T_1) * P(T_2) = 1/2 * 1/2 = 1/4$
1 heads: HT TH 2 heads: HH

Easy Way

1. Find the PMF and write as a table
2. Expand our table by adding columns
3. Add down each column

```
# key is X=x
# value is P(X=x)
# --> points to xP(X=x)
# ---> points to variance
map = {
    0: 0.25 --> 0 ---> 0.25
    1: 0.50 --> 0.5 ---> 0
    2: 0.25 --> 0.5 ---> 0.25
}
```

summation of P(X=x) = 1

summation of xP(X=x) $\mu_x = 1$

summation of variance $sigma^2_x = 0.5$

summation of standard deviation of X ~= 0.7

Example Two

You enter a lottery in which there is a 1 in 1000 chance of winning. If you win, you get $500 and if you don't you get nothing Let Y be the amount of money you win.

Find the PMF, expected value, variance and standard deviation

```
map = {
    0:   0.999 -------> 0 ----> 0
    500: 0.001 ---> 0.5 --> 250 ---> 15.811
}
```

expected value: 0 and 500

Example Two (With a twist)

You enter a lottery in which there is a 1 in 1000 chance of winning. If you win, you get $500 and if you don't you get nothing Let Y be the amount of money you win.

Let V = amount of money you have after the lottery

Find the PMF, expected value, variance and standard deviation

$\Sigma[V] = \Sigma[Y-1] = \Sigma[Y] - 1 = 0.5 - 1 = -0.5$ $Var(V) = Var(Y-1) = Var(Y) = 249.75$ $SD(V) = SD(Y) = 15.8$

Relationship Between Probability and Statistics

Let our random event be:

Pick one person at random and record some characteristics of the individual

The individual we record is the case or unit

The characteristics we record are called variables

The set of all cases of interest: population

Day 4

Outline

1. Statistical Terminology
2. Sampling Distributions

Statistical Terminology

Tidy Data

Each column represents a variable

Header row will contain the name of the variable

Each row represents a case

Each value goes in its own cell.

Good form: left most column contains label variable whose values are unique IDs for the cases

One row could represent:

- a patient
- a particular test for that patient
- all patients seen by a doctor

Merging datasets: you need to pair like data

Data Dictionary

For each variable:

- name of the variable
- type of the variable
- units of measurement
- description

Type of Variable

Numerical (Quantitative): int, float, double

Categorical (Qualitative): string, classes, char, software-specific variable type

Typically, we do not select only one case from the population. We instead select a subset of the population: sample. A sample will always exist in the real world.

Statistical Terminology (Continued)

Variables vary between cases.

Statistics vary between samples

Parameters vary between populations.

Frequentist Statistics

Parameters are constants, but we don't know their value.

Statistics are random variables that describe "randomly select a sample of some fixed size, record values of a variable for each case in the sample, and summarize the value".

In this class

Numerical variables: we use μ to represent population mean and $\bar{X}$ to represent sample mean

Categorical variables: we use "p" to represent population proportion of outcome in a particular category.

We use $\hat{p}$ to represent sample proportion of outcomes in a particular category.

Example

A clinical trial compares two bladder cancer drugs:

- Drug A (Company's "new" drug)
- Drug B (Current best drug)

They recruit 200 subjects with bladder cancer and assign 100 to take Drug A and 100 to take Drug B

Questions

- What is the case
- We can consider this study to have 2 "hypothetical" populations. What are they?
- What are the two samples from those hypothetical populations? (subset of a larger group/population)
- Name an outcome the drug company might be interested in. Is that outcome a numerical or categorical variable?
- What statistic might we use to summarize that outcome in a sample
- What is the corresponding parameter in the hypothetical population.

Answers

- A case is one patient with bladder cancer
- Everyone on Drug A (all bladder cancer patients, if they took Drug A) and everyone on Drug B
- 100 people who took Drug A and 100 people who took Drug B
- How much more effective is Drug A compared to Drug B. This would yield a numerical value. (Reduction in tumor size, cancer remission/not in remission)
- Sample mean ($\bar{X}$) tumor reduction. Sample proportion/sample percent of patients who are in remission.
- Population mean tumor reduction and population proportion in remission

Sampling Distribution

These are things that do not have real world equivalences.

The probability distribution of a statistic is its sampling distribution

Distribution of a statistic over all possible samples of a given size from the sample population. **Must** specify size of sample.

To find a sampling distribution:

1. Simulation: approximate the sampling distribution by simulating samples.
2. Asymptotic behavior: as the number of samples $\to \infty$, what does the distribution look like?

Properties of Sampling Distributions

Let X be the statistic we use to estimate a parameter θ for a population. X is an unbiased estimator of θ if $\mu_{\text{x}} = \theta$. Otherwise X is biased and the amount of **bias** is $\mu_{\text{x}} - \theta$.

The variability of X describes the amount by which individual realizations of X are "spread" about μ_{x}.

We can summarize variability by variance, standard deviation, standard error, margin of error

Day 5

Outline

1. Sampling Distributions
2. Binomial Setting and Sampling Distribution

Sampling Distributions

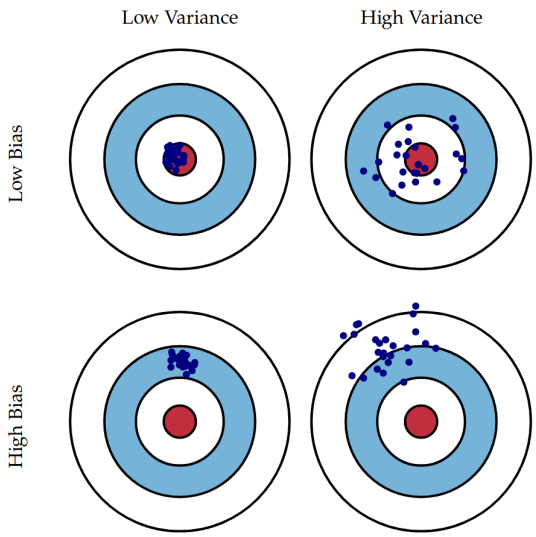


Fig. 1 Graphical illustration of bias and variance.

Figure 1: Illustration of bias and variance

Generally in science we have low bias and high variability.

Joke

A physicist, a biologist and a statistician go hunting:

They are hiding together in the bushes and they see a deer 70ft ahead of them. The physicist makes some calculations, aims and fires at the deer. His shot ends up 5ft to the left of the deer. The biologist analyzes the deer's movement, aims and fires. His shot ends up 5ft to the right of the deer. The statistician drops his rifle and happily shouts, "WE GOT IT!!"

In MATH-338, we assume our sample is generated using random sampling methods (simple random sampling)

- All samples of size `N` are equally likely

If we do not use good sampling methods:

- Probability distribution of sample changes $\rightarrow$ introduces bias

Sampling distribution depends on statistic & sample size

- For very large populations ($\sim$ 20x sample size)
- Sampling distribution does NOT depend on population size

Variability of sampling size $\downarrow$ as the sample size $\uparrow$.

Bias of sampling distribution $<<$ Bias introduced by bad sampling/study design

Variability of sampling distribution $<<$ Variability due to bad sampling/study design

Binomial (Probability) Distribution

Describe the number of "success" in `N` trials in the binomial setting

Four Conditions

Binary outcomes:

- All outcomes are classified either success or failure

Independent outcomes: (hand waved by good study design)

- The previous outcome does not influence the next outcome (flipping a coin, gender of a baby).

Number of outcomes = `N`

- Fixed sample size/number of trials
- Known in advance before any outcomes observed

Success is equally likely for each case/on each trial

- P = Probability of success = population probability of success
- the pass/fail rate of a class is 90/10, everyone is equally likely to fit these odds

Situations

- N term is violated : World Series games. A minimum four games is played but there is no fixed amount of games played
- I term is violated : teams who play away games vs home games

Binomial Random Variable

Let X = count(number) of success in a set of N outcomes obtained in the binomial setting

X has a PMF defined by:

$P(X = x) = \binom{n}{x} P^{x}(1 - P)^{n-x}$

- $n(x) = \frac{n!}{x!(n-x)!}$
  - Number of ways to get x success and n-x failures out of N trials
- probability of getting exactly x success
- Probability of getting n-x failures

Shorthand to: $X \sim B(n, p)$

If you didn't understand them, an extreme simplification would be to say that you are repeating an activity with a chance of success p. Each repetition has the same chance of happening. This chance cannot be affected by the results of the other repetitions. You do this n times. X would represent the number of successes you got after doing n repetitions.

For example, if you toss a fair coin 10 times, the number of heads you will get is:

$$X \sim B(10, 0.5)$$

because you toss it 10 times, and each toss has a 50% = 0.5 chance of being a head (because it is a fair coin).

Figure 2: Explanation

Example One

Toss a fair coin 8 times and let X = number of heads

Is X a binomial random variable?

- B: ✓success = heads
- I: ✓
- N: ✓n=8
- S: ✓p=0.5

$X \sim B(8, 0.5)$

Mean and Variance of a Binomial Random Variable

Consider the Bernoulli random variable:

```
X = {
    1: if outcome is success
    0: if outcome is failure
}
```

If p = probability of success, then:

$\mu_x = 1 * P + 0(1 - P) = P$

Binomial random variable is sum of `N` independent Bernoulli random variables

So if mean of binomial random variable = $P + P + P... + P = nP$ (n number of times)

Variance of Bernoulli random variable =

$(1 - P)^2 P + (0 - P)^2 (1 - P) = P(1 - P)$

When X = 1 and X = 0 respectively ↑

$\Sigma(X - \mu)^2 P(X = x)$ $(1 - \mu)^2 P(X = 1)$

Variance of binomial random variable =

$P(1 - P) + P(1 - P) + ... + P(1 - P) = nP(1 - P)$ (n number of times)

Standard deviation if binomial random variable = $\sqrt{nP(1 - P)}$

Example Two (Question)

Sample of 2000 men get Gemfibrazil

Sample of 2000 men get some other drug (placebo)

It was assumed that 4% of men of the placebo group age get heart attacks (without drug intervention)

Looking at just placebo group:

- Define success, failure, `N`, `P` in this binomial setting
- Find mean, variance, and standard deviation of number of heart attacks the placebo group would experience.

Example Two (Answers)

Formulas used from section: Mean and Variance of a Binomial Random Variable

1

- Success is someone getting a heart attack
- Failure is someone not getting a heart attack
- 2000 is `N`
- 0.04 is `P`

2

- Mean: 80
- Variance: 76.8
- Standard deviation: 8.76356092

Distribution of Sample Proportion

$\hat{P} = \frac{X}{n}$

We have no idea how to estimate the number of successes in a very large population

Proportions are restricted to $[0, 1]$

Sample proportion and population proportion are on the same scale

$E[\hat{P}] = E[\frac{X}{n*X}] = \frac{X}{n} * E[X]$

If $X \sim B(n, P)$ then $E[X] = \mu_x = nP$

$E[\hat{P}] = \frac{X}{n}(n - P) = P$

$\hat{P}$ is an unbiased estimator of P

Variance$(\hat{P}) = V(\frac{1}{n} * X) = (\frac{1}{n})^2 V(x)$

If $X \sim B(n, P)$ then $V(x) = nP(1 - P)$

$V(\hat{P}) = (\frac{1}{n})^2(nP(1 - P)) = \frac{P(1-P)}{n}$

$SD(\hat{P}) = \sqrt{\frac{P(1-P)}{n}}$

Probability Problems involving p hat

$\hat{P}$ does NOT have a binomial distribution

However we can convert $X = n * \hat{P}$ and X has a binomial distribution

Day 6

Outline

1. Types of Studies
2. Design Experiments
3. Why studies go wrong

Types of Studies

Observational

We simply observe things and do not manipulate them

Figure 1: observing

Experimental

Manipulate one or more explanatory variables and record response variables

Try to prove cause and effect

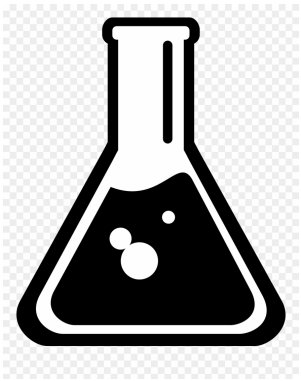


Figure 2: experimenting

Design of Experiments: Terminology

Manipulated explanatory variables: experimental variables, factors

- Treat as a categorical variables
- Values are called levels of the factor

A set of conditions caused by combining levels of different factors: treatment

Recorded response variables: outcomes

Example One

Experiment to compare two weight-loss drugs

Two separate factors

- Drug (A & B)
- Diet (Normal & Special)

Full factorial experiment: 4 treatments

- Drug A & normal diet
- Drug B & normal diet
- Drug A & special diet
- Drug B & special diet

Example Two

Visitors to a website.

Two separate factors

- Placement of an ad (top/bottom)
- Background color (blue/red)

Full factorial experiment: 4 treatments

- Top & Blue
- Bottom & Blue
- Top & Red
- Bottom & Red

Interacting Variables

Factors can be interacting variables effect of factor one on response changes depending on value on factor two.

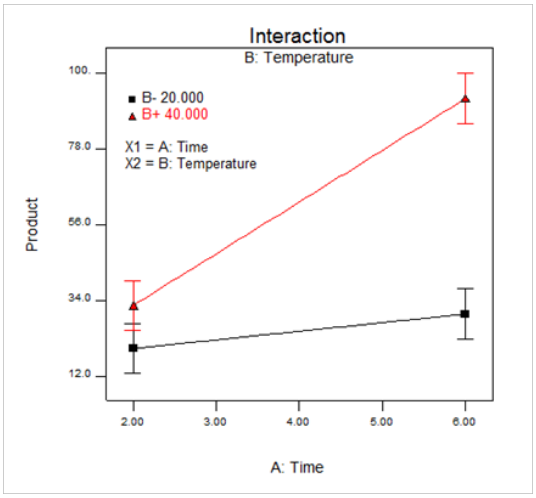


Figure 3: Interaction effect

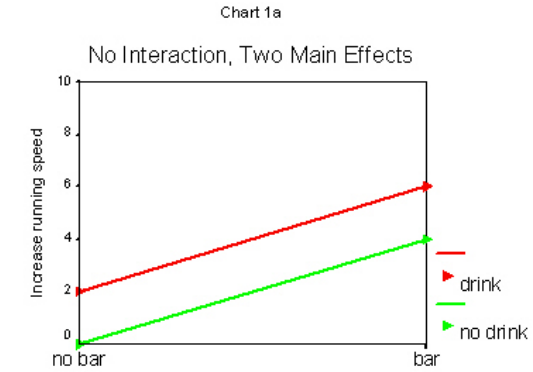


Figure 4: No interaction effect

Confounding Variable

Two variables (factors or not manipulated explanatory variables) are confounding variables if their effects on response cannot be disentangled/distinguished.

Often but not always, confounding variables are related to each other.

Lurking variables are potential confounding variables that we ignore during study design.

Formal definition: In statistics, a confounder is a variable that influences both the dependent variable and independent variable, causing a spurious association. Confounding is a causal concept, and as such, cannot be described in terms of correlations or associations.

Examples

- Age

Three Principles of Experimental Design

Control

Subjects on each treatment should be as "similar" as possible

"controlling" for effects of other variables than our experimental variable

Randomization

Treatments should be randomly assigned to subjects

Replication/Repetition

Other people can repeat your experiment on similar subjects and get similar results

We have enough subjects to "eliminate" variability issues

Types of Experimental Designs

Completely Random Design

Subjects assigned a treatment entirely at random

Block Design

Non randomly divide our subjects into groups ("blocks") based on potential confounding variables, then assign treatments within each block.

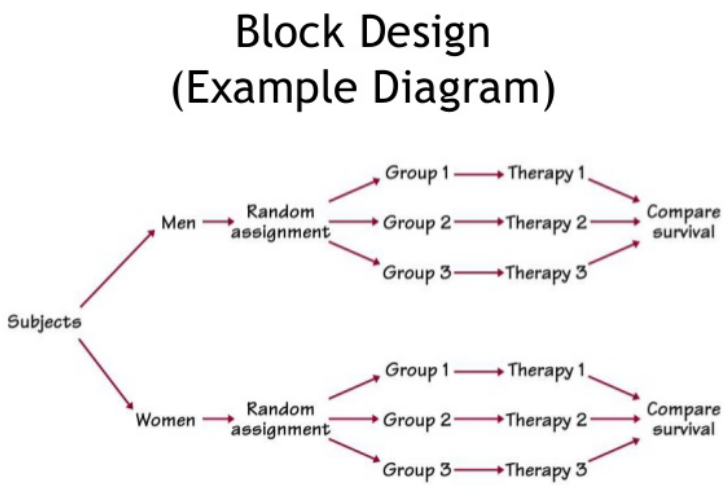


Figure 5: Block Design

Matched Pairs Design (Paired Design)

Special case of block design with blocks size of two (only looking at one treatment with two levels)

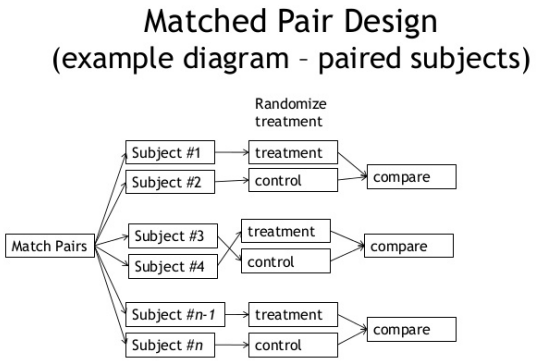


Figure 6: Matched Pair Design

Repeated measures design

More than two levels

Idea: get you & someone as similar to you as possible.

Randomly assign you to one treatment and assign your "pair" to other treatment

Often, we have same subjects undergo all treatments or record response at multiple times.

These are matched pairs/repeated measures designs in which order of treatments is manipulated/randomized.

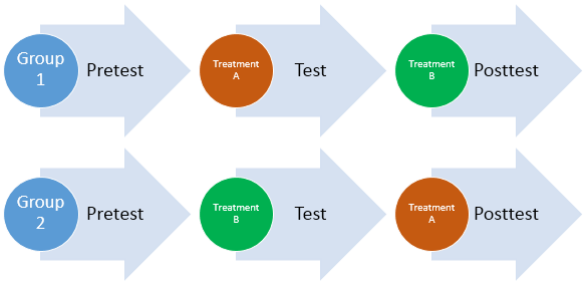


Figure 7: Repeated Measure Designs

One of the levels of treatment is control level

Group whose treatment entirely control levels = control group

In this class: one control group ≥ 1 "treatment group"

Why Studies Go Wrong

1. You accidentally introduce a confounding variable
    - "Placebo effect": Otherwise ineffective treatment "works" because people think it will
    - Effects due to people involved in experiment knowing who is in each group
    - Blinding/masking: not letting experimenters or subject know which group anyone is in
    - Double blind: neither subject nor experimenter know
2. Equipment failure/mistake in recording data
3. Hawthorne Effect
    - If people know they're in the study they change their behavior
4. Subjects are not representative of the target population
    - Using rats for experiments where rats did not respond the way they needed to
5. Treatments are not representative of real world conditions

Day 7

Outline

1. Two-way tables and diagnostic testing
2. Conditional probability
3. Tree Diagrams
4. Bayes' Rule
5. Solving conditional probability problems

Two-Way Table

Gender compared to handedness

	Handed		
	Left	Right	
Female	7	46	53
Male	5	63	68
	12	109	121

Figure 1: Table Example

Example

Helsinki Heart Study

2035 men in control group had 84 heart attacks 2046 men in special drug name group had 54 heart attacks

Referring to the table above, these are the correct mappings:

- Male: placebo
- Female: special drug
- Left: Heart attack
- Right: No heart attack

Diagnostic Testing

Formal definition: an examination to identify an individual's specific areas of weakness and strength in order determine a condition, disease or illness.

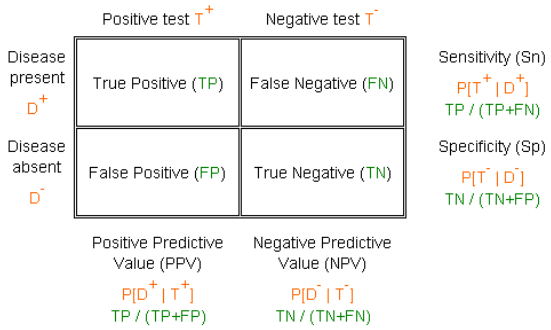


Figure 2: Diagnostic Testing Table

Machine Learning Terminology: we evaluate on a "training set" in which number if actual positive and actual negative is known in advance.

We compute:

- Sensitivity: proportion of actual positive classified correctly $\frac{TP}{TP+FN}$.
- Specificity: proportion of actual negative classified correctly $\frac{TN}{TN+FP}$

These are both properties of our test/algorithm.

- Positive Predictive Value(PPV, precision): proportion of positive tests that were actually positive $\frac{TP}{TP+FP}$
- Negative Predictive Value(NPV): proportion of negative tests that are actually negative $\frac{TN}{TN+FN}$

Also depend on prevalence (base rate) $\frac{Actual\ positive}{Actual\ positive+Actual\ negative}$

Example

- 300 units
- 83% prevalence
- TP = 200
- FP = 10
- FN = 50
- TN = 40

Compute:

- Sensitivity: $\frac{200}{200+50} = 80\%$
- Specificity: $\frac{40}{40+10} = 80\%$
- PPV = $\frac{200}{200+10} = 95.20\%$
- NPV = $\frac{40}{40+50} = 44.4\%$

In Class Example

- 300 units
- 3% prevalence
- TP = 8
- FP = 58
- FN = 2
- TN = 232

Answers

- Sens: $\frac{8}{10} = 80\%$
- Spec: $\frac{232}{232+58} = 80\%$
- PPV: $\frac{8}{8+58} = 12.1\%$
- NPV: $\frac{232}{232+2} = 99.2\%$

Conditional Probability

The conditional probability of event "B" given event "A", denote denoted $P(B|A)$, is the probability of event "B", looking only at outcomes in A.

$P(B|A) = \frac{number\ of\ outcomes\ in\ A \cap B}{number\ of\ outcomes\ in\ A}$ when all outcomes are equally likely

More generally: $P(B|A) = \frac{P(A \cap B)}{P(A)} P(A) > 0$

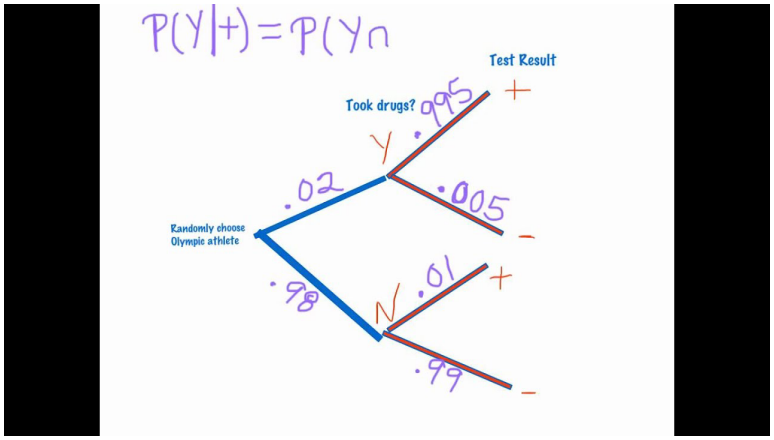


Figure 3: Conditional Probability Diagram

Independent events: $P(A \cap B) = P(A) \times P(B)$

Conditional probability: $P(A \cap B) = P(A) \times P(B|A)$

So: "A" and "B" are independent when $P(B) = P(B|A)$ dependent when $P(B) \neq P(B|A)$

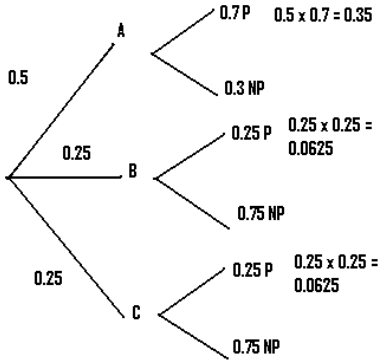
Example

1. What is the probability that a randomly selected person in the control group had a heart attack?
2. What is the probability that a randomly selected heart attack victim was in the control group?

Both of these are different questions

1. $P(heart\ attack|control) = \frac{84}{2035}$
2. $P(control|heart\ attack) = \frac{84}{140}$

Tree Diagram



- ——— = *branch*

Each node represents an event.

Each branch represents probability of getting to next node, given that we got to previous node.

Ending node is the **terminal node**.

Splits in the branches must add up to one. Please refer to node "A" when it splits between "NP" and "P".

Notes on Probabilities

1. Probability of leaving a node is 1. The sum of probabilities on all branches exiting a node = 1.
2. Probability of getting to a terminal node is a product of probabilities along the branch path to it.
3. Probability of event "B" is sum of probabilities at all terminal nodes including "B"

Day 8

Outline

1. A History Lesson
2. Neyman-Pearson Hypothesis Testing

History Lesson

Major Players

- Karl Pearson
- Egon Pearson
- Jerzy Neyman
- Ronald Fisher

TL;DR

This test will allow us to make preemptive decisions based on conditions presented before the study is conducted. These are the theoretical outcomes WITHOUT taking any sample data

Neyman-Pearson Hypothesis Testing

TL;DR version

1. Define a boundary used to inform a decision
2. Obtain data and see which side of the boundary it falls on
3. Make decision

Example

We have a coin and it is weighted but we don't know if it's weighted to be 60% heads or 60% tails.

Define a parameter to describe the situation

Let `P` represent probability of getting heads ("population proportion of heads")

Define two competing "hypothesis" involving the parameter.

(heads)

- H_0 : `P = 0.6` [null hypothesis: "nothing unexpected"]
- H_1 : `P = 0.4` [alternative hypothesis: "something is happening, we should change our minds"]

Define a "critical region" based on our sample data

1. Define a test statistic `T` whose value can be computed from the sample data
2. Define the sampling distribution of `T` under H_0 and H_1
3. Based on the sampling distribution under H_0, define:
   - $\alpha = P(we\,claim\,H_1\,is\,true|H_0\,is\,true)$ and find the region in the sampling distribution under H_0 corresponding to that α value.
4. If the observed value of `T` is in that region, conclude H_1 is true. Otherwise, conclude H_0 is true

Critical region: a range of values that corresponds to the rejection of the null hypothesis at some chosen probability level.

Example

Our decision rule:

- If we get 4 or fewer heads in 10 flips: conclude H_1 is true.
- If more than 4 heads in 10 flips: conclude H_0 is true.

"Critical region": Let $X = number\,of\,heads\,in\,10\,flips$

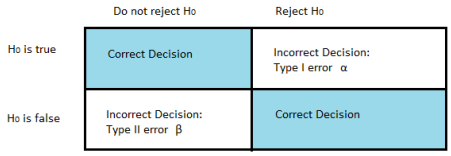
- $X \leq 4$

Recall:

Gender compared to handedness

	Handed		
	Left	Right	
Female	7	46	53
Male	5	63	68
	12	109	121

Now apply this to Neyman-Pearson rules:



Under N-P Rules

Type 1 Error is "worse" than Type 2 Error. However, if P(Type 1 Error) is too low, P(Type 2 Error) balloons.

$\alpha = P(1)$ - P(Concluded H_1 | H_0 is true)

$\beta = P(2)$ - P(Concluded H_0 | H_1 is true)

Power of test $= 1 - \beta$

- = P(concluded H_1 | H_1 is true)

Example [Continued from Above]

Let $X = number\,of\,heads\,in\,10\,flips$

- Under H_0: $X \sim B(10, 0.6)$
- Under H_1: $X \sim B(10, 0.4)$

For critical region $X \leq 4$:

- $\alpha = P(X \leq 4 \,|\, p = 0.6) = 0.166$
- $\beta = P(X > \,|\, p = 0.4) = 0.367$

Power $= P(X \leq 4 \,|\, p = 0.4) = 0.633$

Traditionally, set $\alpha = 0.05$ or $\alpha = 0.01$

- α refers to the probability of making a Type I Error.

Find the critical region giving a Type 1 Error rate of at most α

(Find x such that P(X $\leq$ x | H_0 is true) $\leq \alpha$)

P(x $\leq$ 2 | H_0 is true) = 0.0123

P(X $\leq$ 3 | H_0 is true) = 0.0548

Critical region corresponding to $\alpha = 0.05$: $x \leq 2$

What is β for this critical region?

- $\beta = P(x > 2|p = 0.4) = 0.833$

In most fields, we use power instead

Power $= P(X \leq 2|p = 0.4) = 0.167$

Rules of Thumb

1. $\alpha < \beta$. If $\alpha \leq \beta$, either decrease α or switch H_0 or H_1
2. At your "given" α value, $\beta \leq 2$ or equivalently, power ≥ 0.8 (80% power). If power < 0.8, plan to collect more data!

Power must be at least 80 percent

In Practice

1. The idea of "nothing weird happening" should give us the value of the parameter.
2. We define a clinically signifiant/practically signifiant difference in parameter values ("minimum effect size")

What we need at each step

1. To compute the critical region:

- need α, H_0 (value of P under H_0)
- sampling distribution of test statistic under H_0

2. To compute power:

- need critical region, H_1 (value of P under H_1)
- sampling distribution of test statistic under H_1

Day 9

Outline

1. Conditional probability example
2. Power analysis example

Please reference the attached sheets for full examples

Conditional Probability Examples

Example 1

In a lecture class of 150 students, 110 students are freshmen, 50 own a dog, and 25 are freshmen who own a dog. Suppose a student is selected at random.

Tree Diagram Version

Root

- Freshman ($\frac{110}{150}$)
  - Own dog : $\frac{25}{110}$ [Freshman AND own dog = ($\frac{110}{150} \times \frac{25}{110} = \frac{1}{6}$)]
  - No dog : $\frac{85}{110}$ [Freshman AND no dog = ($\frac{110}{150} \times \frac{85}{110} = \frac{17}{30}$)]
- Not Freshman ($\frac{40}{150}$)
  - Own dog : $\frac{25}{40}$ [Not freshman AND own a dog = ($\frac{40}{150} \times \frac{25}{40} = \frac{1}{6}$)]
  - No dog : $\frac{15}{40}$ [Not freshman AND not own a dog = ($\frac{40}{150} \times \frac{15}{40} = \frac{1}{10}$)]

a. What is the probability of being a freshman, given that the student owns a dog?

- $P(Freshman|Dog) = \frac{P(Freshman\ AND\ Dog)}{P(Dog)} = \frac{\frac{25}{150}}{\frac{25}{150}+\frac{25}{150}} = \frac{25}{50} = \frac{1}{2}$

b. What is the probability of owning a dog, given that the student is a freshman?

- $P(Dog|Freshman) = \frac{P(Dog\ AND\ Freshman)}{P(Freshman)} = \frac{\frac{25}{150}}{\frac{110}{150}} = \frac{25}{110} = \frac{5}{22}$

- P(Freshman & Dog) = P(Freshman) P(Dog|Freshman)
- P(Freshman & Dog) = P(Freshman) P(Dog) ← Independence

Table Diagram Version

	Freshman	Not Fresh.	Total
Dog	25	25	50
No Dog	85	15	100
Total	110	40	150

Figure 1: Freshman Table Example

Power Analysis Examples

Example 1

It is believed that about 10% of the population is left-handed. However, China has claimed that less than one percent of its students are left-handed. Suppose we are interested in evaluating whether there is something special about Chinese people, or whether the Chinese government is lying. Suppose further that we have devised a scientifically perfect test to measure a person's dominant hand. Would a random sample of 50 Chinese students be large enough to detect a population difference of 10% vs. 1%?

We want low α and high power

- H_0 : p = 0.1 [Null]
- H_1 : p = 0.01 [Alternate]
- `N` : 50
- α: 0.05
  - If α is not given, please assume $\alpha = 0.05$
- Define p = proportion of left handed students
- For midterm one, define X = number of (successes) left-handed students in our sample.
- Decision rule:
  - Critical region: $X \leq x$
  - If X is in critical region, accept H_0, else accept H_1.
  - Only problem is we do not know what `x` is.
  - Defining our critical region to be $X \leq 5$
    - Under the null hypothesis H_0, $X \sim B(50, 0.1)$
      - $P(X \leq 5 | p = 0.1) = 0.616$
      - P(Type 1 Error) = 0.616
    - Under the alternative hypothesis H_1, $X \sim B(50, 0.01)$
      - $\beta = P(X > 5 | p = 0.01) = 0$
      - Power $= P(X \leq 5 | p = 0.01) = 1$
  - Defining our critical region to be $X \leq 1$
    - When p = 0.1
      - 3.4% false positive
      - 96.6% true negative
      - $\alpha = 0.034$
    - When p = 0.01
      - 91.1% true positive
      - 8.9% false negative
      - Power = 0.911
      - $\beta = 0.089$

Example 2

Is this sample large enough to detect something → power rule!!!!!!!

We want low α and high power

- H_0 : p = 0.26 [Null]
- H_1 : p = 0.52 [Alternate]
- `N` : 14
- α: 0.05
  - If α is not given, please assume $\alpha = 0.05$
- Define p = % of patients progression free after 6 months

Using `R`:

- Critical region is X > 6 or $X \geq 6$
- `lower.tail = TRUE` includes $\leq$
- `lower.tail = FALSE` includes >
- When P = 0.26
  - 4.7% false positive
  - 95.3% true negative
  - $\alpha = 0.047$
- When P = 0.52
  - 66.2% true positive
  - 33.8% false negative
  - Power = 0.662
  - $\beta = 0.338$

Day 10

Outline

1. “Null Hypothesis Significance Testing”
2. When Null Hypothesis Significance Testing goes horribly wrong

Null Hypothesis Significance Testing

Recall

- identifying a parameter is not "too hard"
- identify its value under H_0 is trivial
- However, identifying its value under H_a is difficult in practice
- Under N-P (Neyman-Pearson): define minimum effect size
- But often, we have no idea
- This is what we have been doing for the past 70 years and it does not require any subject knowledge. "It just works". Will also allow us to get the `P` value.

NHST: just give the inequality in alternative hypothesis H_a

Suppose H_0: $\theta = \theta_0$

- $\theta \rightarrow$ arbitrary parameter
- $\theta_0 \rightarrow$ its value under H_0

N-P:

- H_1: $\theta = \theta_1$
- $\theta_1 \rightarrow$ its value under H_1

NHST: Choose from

- H_a: $\theta > \theta_0$
  - Theory says θ should be bigger
  - One-tailed, one-sided hypothesis testing
    - One-tailed testing: The critical area of a distribution is either $<$ or $>$ a certain value but not both
- H_a: $\theta < \theta_0$
  - Theory says θ should be smaller
  - One-tailed, one-sided hypothesis testing
- H_a: $\theta \neq \theta_0$
  - No idea what to expect or theory suggests arguments for both $>$ **AND** $<$
  - Two-tailed, two-sided hypothesis testing
    - Two-tailed the sample is greater than or less than a certain range of values:

NOTE: method of collecting data tells us what θ is (what θ_0 is) and may suggest H_a

When in doubt: use the H_a version with $\neq$

Next Step : Distribution

Define a test statistic whose value will be computed from sample data

N-P: Find its distribution under both H_0 & H_1

NHST: Find its distribution under H_0 but we don't know its distribution under H_a

Next Step : Critical Region

Define a critical region of the test statistic such that if the observed value is in critical region, accept H_1

NHST: Define a critical region such that if in critical region, reject H_0.

If not in critical region, fail to reject H_0

Example (Theory) : Jury

Start off assuming innocence (H_0) [Null Hypothesis]

- Prosecution presents evidence (test statistic observed value)
- Jury decides if it enough evidence
  - Enough evidence (in critical region) $\rightarrow$ reject the assumption of innocence and declare guilty (reject H_0 and acccept H_a)
  - Not enough evidence (not in critical region) $\rightarrow$ fail to reject presumptions of innocence. He might still be guilty but the evidence is not damming enough to convince us otherwise. We fail to reject H_0 or the Null Hypothesis

In NHST, define significance level (not α but works like α). Here, the significance level is the probability of rejecting the null hypothesis which is generally the same value of α

This can go horribly wrong because power is not taken into account also, there is not a big enough sample size to correctly draw a conclusion.

P-Value

A measure of the "strength" of the evidence against H_0. This is related to power?

ALWAYS COMPUTED AFTER OBSERVATION

Official definition: probability of obtaining our observed value of the test statistic, or a value as or more favorable to H_a, if H_0 is true.

- $P(X \geq x_{observed} \mid H_0 \text{ is true})$ when H_a: $\theta > \theta_0$
- $P(X \leq x_{observed} \mid H_a : \theta < \theta_0)$

Things go weird for two-tailed tests

Usually: P(X is equally or less likely than $x_{observed} \mid H_0$ is true) but sometimes we get one-tailed p-values

"How likely is it that I got this lucky or luckier?"

When P-Value $\leq$ significance level

1. H_0 is true and I got really lucky $\leftarrow$ I will make a Type I Error
   - H_0 may in fact be false but you have circumstantial evidence to promote the notion that H_0 is true
2. H_0 is not true and H_a
   - This is the ideal situation as we do not make any false assumptions about H_0

Either way, I reject H_0 and conclude H_a is true

When P-Value $>$ significance level

1. H_0 is true
2. H_0 is not true, but we don't have "unlikely enough" evidence
   - In this case, we acquit an guilty man. Even though we damn well know he did it but the prosecution did not provide damning evidence to conclude that he is guilty.

Either way, we fail to reject H_0 (make no conclusion so default to assumption H_0 is true)

Example (Book Exercise 8.20 [Application])

Study of children, program intended to increase consumption of whole grains. At end of program, sample of 86 children got a snack.

- 48 children chose whole grain
- 38 chose regular

Suppose that before program, children were equally likely to pick either snack. Do we have enough evidence to claim the program works as intended?

Step 1 (Identify Parameter of Interest)

Use it to write H_0 and H_a.

- `p` = Proportion of all children who choose whole grain (generalized results)
- H_0 : `p = 0.5`
- H_a : `p > 0.5`

Step 2 (Identify Test Statistic and its sampling distribution under H_0)

Let `X` = number of success (number of children in sample choosing whole grain)

$X \sim B(n = 86\,, p = 0.5)$

Step 3 (Observe data and calculate value of test statistic)

$X_{\text{observed}} = 48$

Step 4 (Calculate EITHER the critical region or the P-Value)

P-Value is way easier to compute when you have software

```
binom.test(x = 48, n = 86, p = 0.5, alternative = "g")
```

From software: p-value = 0.166

Step 5 (Determine whether or not to reject H_0)

5% is our cut-off. If the value is LESS than 5% then we **reject the Null Hypothesis.**

Since $0.166 > 0.05$, our results are likely enough under H_0 therefore, we fail to reject H_0

Step 6 (Write what "reject H_0" or "fail to reject H_0" means in context)

We do not have "statistically signifiant" evidence to claim that the program is working. It is reasonable to continue with the assumption that children are still equally likely to pick a healthy snack. We failed to reject the notion that they will be more likely to pick a healthy snack.

When Null Hypothesis Significance Testing goes horribly wrong

1. Very small samples
2. Very large samples
   - Neyman-Pearson: $p = 0.5$ vs $p = 0.50001$
   - NHST: $p = 0.5$ vs $p > 0.5$
3. Significance level is not arbiter of importance (2 and 3 are practically the same)
4. Lots of tests
5. P-hacking

Day 11

Outline

1. Fisher's Significance Tests
2. Goodness of fit test

Tl;DR

This model is concerned about the model used rather than actually trying to prove anything. If the initial hypothesis is rejected then we look into why it failed, rather than stopping at that conclusion.

Midterm Exam 1

Lecture

- ~ 60%
- 4 to 5 multi-part problems

Allowed: one-sided formula sheet (can be typed or printed)

Lab

- ~ 40%
- 3 problems, 2 to 3 part problems

Allowed: textbook, notes, software help, anything on Titanium

Fisher's Significance Testing

In Fisher's view, there is only one hypothesis. We see how well sample data "fit" that hypothesis.

In the strictest sense, the hypothesis includes all assumptions about the probability model used to obtain the sampling distribution of the test statistic.

Assumption are of two kinds:

1. Assumptions about parameters
2. Assumptions about data generation/collection

In practice, we refer to the hypothesis as the null hypothesis (H_0).

Using the Null Hypothesis Significance Testing we write H_a: not H_0

Where we see this:

- χ^2 test (Chi-Squared)
- ANOVA

In Fisher's approach, we first specify our model

- Then: Specify sampling distribution of test statistic
- Then: Collect data and compute value of test statistic
- Then: Get P-Value

Recall: P-Value is probability of obtaining our data, or a result with a test statistic signalling equal or greater "distance" from H_0, if H_0 is true.

Pure Fisher Philosophy: Stop here. P-Value represents "how well" data fit hypothesis.

- Very high P-Value: data fit suspiciously well
- Very low P-Value: data does not fit well at all

In practice: Fisher commends a personal significance level.

- Significance = "Signifying something"

P-Value $\leq$ significance level: our results are a meaningful difference from the model. We should investigate!

- Reject H_0

P-Value $>$ significance level: our results are consistent with the model. We did not prove it correct but the model is a reasonable approximation of reality.

- Fail to reject H_0

In practice: we define one main assumption about the parameters to be the H_0 that can get rejected.

Goodness of Fit Testing

Most of the "classic" goodness of fit tests involve genetics.

Example (Theory) : Mendel's Pea Plants

Dihybrid cross for seed shape & seed color Mendel's Laws: Should see a 9:3:3:1 ratio

Hypothesis: $\frac{9}{16}$ round/yellow, $\frac{3}{16}$ round/green, $\frac{3}{16}$ wrinkled/yellow, $\frac{1}{16}$ wrinkled/green

In practice:

H_0: $P_{RY} = \frac{9}{16}$, $P_{RG} = \frac{3}{16}$, $P_{WY} = \frac{3}{16}$, $P_{WG} = \frac{1}{16}$

We know the observed sample will probably not have those proportions.

We need a measure of "how far off" our sample is from what we expect. [χ^2 test statistic]

We define (Pearson) residuals for the different categories:

Residual = $\frac{O-E}{\sqrt{E}}$

- O : # of observed in sample
- E : # of expected in sample

$\chi^2 = \Sigma\ residual^2 = \Sigma\ \frac{O-E}{\sqrt{E}}$

Where Σ is all categories

Mendel observed:

- 315 RY
- 108 RG
- 101 WY
- 32 WG

Total: 556 seeds

What do we expect?

- $\frac{9}{16}(556) = 312.75$ RY
- $\frac{3}{16}(556) = 104.25$ RG
- $\frac{3}{16}(556) = 104.25$ WY
- $\frac{1}{16}(556) = 34.75$ WG

Pearson Residuals:

- RY : $\frac{315-312.75}{\sqrt{312.75}} = 0.127$
- RG : $\frac{108-104.25}{\sqrt{104.25}} = 0.367$
- WY : $\frac{101-104.25}{\sqrt{104.25}} = -0.318$
- WG : $\frac{32-34.75}{\sqrt{34.75}} = -0.467$

Goodness of Fit Testing (Continued)

Contribution of a category χ^2 = square of its Pearson residual

In our example:

$\chi^2 = (0.127)^2 + (0.367)^2 + (-0.318)^2 + (-0.467)^2 = 0.47$

Our sample data is 0.47 "off" from what we expected.

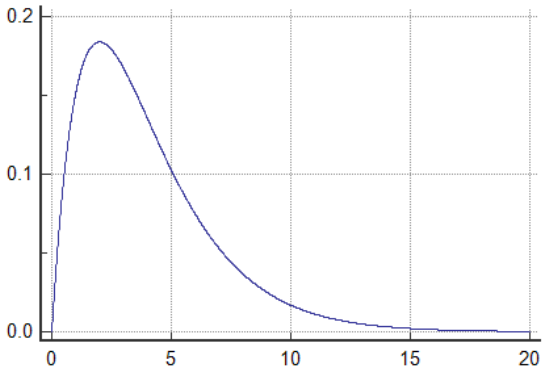
χ is unit-less.

To compute the P-Value, we need a sampling distribution of χ^2

P-Value = P($\chi^2 \geq 0.47$ | H_0 is true)

Approximate the sampling distribution 1 of 2 ways:

1. Under H_0, χ^2 has approximately a χ^2 distribution with (# of categories - 1) degrees of freedom



- Strictly non-negative
- "Skewed right"

Software gives us approximate-value $P(\chi^2 \geq 0.47) = 0.9254$

If P-Value is "really small": reject H_0 means "our proportions are not all correct" - we should investigate to find out which ones & why.

Mendel's data was probably full crap, not on him but the guy collecting the peas.

2. Simulate very many samples of sizes `n`, under assumption of H_0 is true, and compute χ^2 for each simulated sample

When we expected ≥ 5 in each category in our sample, both approaches give similar results.

When in any category we expect < 5 counts, we use #2 above.

Day 12

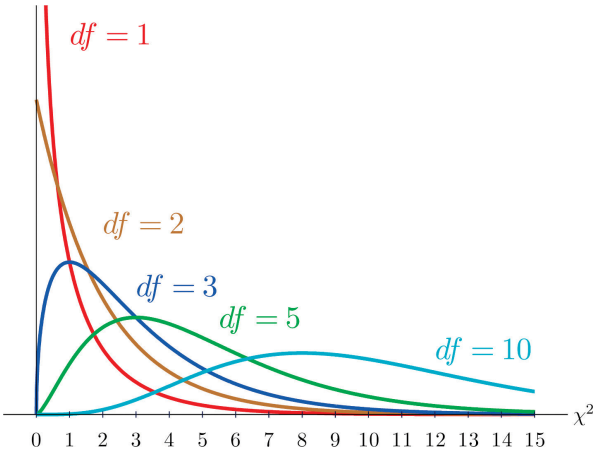
Outline

1. Computing p-values with software using χ^2 distraction
2. Test of independence
3. Measures of association between two categorical variables

Computing P-Value

Given χ^2

P-Value is always above or equal to degrees of freedom.



p ___

- interval values → probability

q ___

- probability → values

P-Values = probability of getting our data or data that "disagree" as much or more with our model, if model is correct.

```
arbitrary_val <- 4.8
pchisq(arbitrary_val, df = 5, lower.tail = FALSE)
```

Test of Independence

What I'm recording: two categorical variables

What I want to know: whether a suspected association between the variables will hold when generalized to the population.

Test of Homogeneity

What I'm recording: 1 categorical variable in samples form multiple populations

What I want to know: Is the variable's distribution the same in all populations?

Both tests use data summarised in two-way tables

We use Fisher's significance testing approach.

Test of independence: we model assuming that the two variables are not actually associated

H_0

- [Variable 1] does not affect [Variable 2]
  - [Variable 1] and [Variable 2] are independent/not associated/ not related

Testing of homogeneity: we model assuming the distribution is the same in every population

- H_0: the distribution of [variable] is the same in [list of population]

More simply put: H_a: not H_0

In test of homogeneity, we consider "population" to be an explanatory variable & run a test of independence

Observed counts = number in sample of each cell of table.

Example (Book Example 9.12)

	Low Salt	High Salt	Total
CVD			200
NO CVD			2215
Total	1169	1246	2415

Figure 1: Base Table

Estimated probability of Cardiovascular Disease(CVD) = $\frac{200}{2415}$

If independent (according to chart):

- P(CVD | low salt) = $1169 \times \frac{200}{2415} = 96.81$
- P(CVD | high salt) = $1246 \times \frac{200}{2415} = 103.19$
- P(NO CVD | low salt) = $1169 \times \frac{2215}{2415} = 1072.19$
- P(NO CVD | high salt) = $1246 \times \frac{2215}{2415} = 1142.81$

Pearson Residuals

$\frac{O-E}{\sqrt{E}} \rightarrow$ for each cell

Contribution of a cell to χ^2: residual^2 $= \frac{(O-E)^2}{E}$

$\chi^2 = \Sigma \frac{(O-E)^2}{E}$

- P(CVD | low salt) $= \frac{88-96.81}{\sqrt{96.81}} = -0.895$
- P(CVD | high salt) $= \frac{112-103.19}{\sqrt{103.19}} = 0.867$
- P(NO CVD | low salt) $= \frac{1081-1072.19}{\sqrt{1072.19}} = 0.269$
- P(NO CVD | high salt) $= \frac{1134-1142.81}{\sqrt{1142.81}} = -0.261$

$\chi^2 = (-0.895)^2 + (0.867)^2 + (0.269)^2 + (-0.261)^2 = 1.69$

finish second chart from picture

To get a P-Value

- Option 1: Our χ^2_{observed} value comes from a χ^2 distribution with degrees of freedom. Find $P(\chi^2 \geq \chi^2_{\text{observed}})$
- Option 2: Simulate a bunch of samples assuming independence, then find proportion of simulated χ^2 statistic $\geq \chi^2_{\text{observed}}$

Fisher: df (degrees of freedom) = (r - 1)(c -1)

- r: rows
- c: columns

"Sample size assumptions" method (2) always works but different people can get different values.

Method 1 always gives some value, but that value can be inaccurate at small sample sizes.

When all expected counts ≥ 5, use method 1.

When any expected count < 5, use method 2.

Alternate method when `n` is really small: Fisher's exact test

Condition on marginal totals being fixed, get a test statistic with hypergeometric distribution.

P-Value = $P(\chi^2 \geq 1.69)$ from χ^2 distribution with 1 degree of freedom = 0.193

Not on test but may show up in context:

3 "Measures of association" between categorical variables

1. Difference in proportions

- Population: P_1 - P_2
- Samples: $\hat{P}_1 - \hat{P}_2$

2. Relative risk (RR)

- Population : $\frac{P_1}{P_2}$

Day 13 (Review Session)

Problem Solving Methodology

1. What problems look like
2. What info to look for
3. How to solve them

Independent vs. Disjoint

Are A and B Disjoint

Looks like: define two events

Looking for:

- P(A)
- P(B)
- P(A ∩ B)

Check: Is P(A ∩ B) = 0

If yes → disjoint

If no → not disjoint

Are A and B independent

Looking for:

- P(A)
- P(B)
- P(A ∩ B)

Find ONE OF the following comparisons:

Is P(A ∩ B) = P(A)P(B)

If yes → independent

If no → dependent

Is P(B | A) = P(B)?

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

Event vs Probability

- **Event:** an actual thing that could happen or not
  - Example: the dog will go to the right place
- **Probability:** a number between 0-1 assigned to the "chance" of an event
  - Example: The dog has a 50% chance of going to the right bowl

Example

Problem: 3 lights, independently red/green

If red $\rightarrow$ 2 minutes

10 minutes to get to work

All green $\rightarrow$ 8 minutes

P(A) = 0.6 P(B) = 0.4 P(C) = 0.9

A $\rightarrow$ Light 1 is red B $\rightarrow$ Light 2 is red C $\rightarrow$ Light 3 is red

Goal: Find the probability of not being late to work

Employee is not late if:

- all green
- one light is red

P(not late) = P(all green) + P(one red)

all green = $\{(A^c, B^c, C^c)\}$ **he hits no red lights**

one red = $\{(A, B^c, C^c)$, (A^c, B, C^c), $(A^c, B^c, C)\}$ **he hits at most one light**

all green = P($\{(A^c, B^c, C^c)\}$)

= P(A^c)P(B^c)P(C^c)

= (0.4)(0.6)(0.1)

= 0.024

finish from picture

Conditional Probability/Bayes' Rule

Given: two conditional probabilities/proportions $P(B|A), P(B|A^c)$

One unconditional probability/proportions (prevalence/base rate) P(A)

Goal: Find a different conditional probability P(A|B)

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

Contents

Day 14

Continuous Random Variables

Can take any real number ($\mathbb{R}$) value within any given interval.

We cannot use a probability mass function so we will instead use a probability density function (PDF) denoted as $f(x)$

Properties

- The probability of being in an interval (a, b] is:

$$\int_a^b f(x)dx = \int_{-\infty}^b f(x) - \int_{-\infty}^a f(x)dx$$

  - This is considered the area under the curve between a and b
- $P(X = x) = 0 \ \forall x$
  - $P(X \leq x) = P(X < x)$
  - $P(X \geq x) = P(X > x)$

$f(x)$ is displayed graphically as a density curve

Properties of $f(x)$

- $\forall \ x \in \mathbb{R}$, $f(x) \geq 0$
  - Density never goes below x-axis
- $\int_{-\infty}^{\infty} f(x)dx = 1$

Mean of continous random variable: $\mu_x = \int_{-\infty}^{\infty} x \times f(x)dx$

Variance of continous random variable is $\sigma^2 = \int_{-\infty}^{\infty} (X = \mu_x)^2 f(x)dx$

Uniform Random Variable

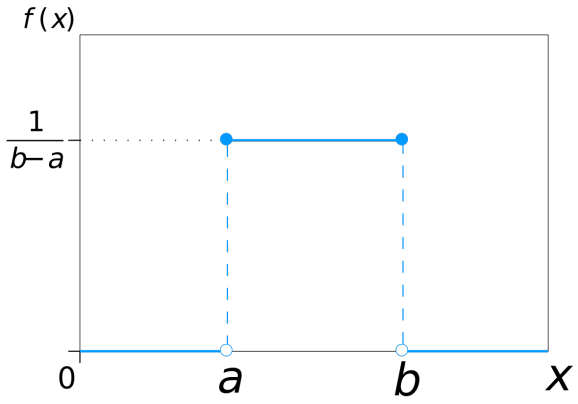


Figure 1: Graphical Representation

$X \sim U(a, b)$

Example : Standard Uniform Random Variable

$X \sim U(0, 1)$

Find

- P(X $\geq$ 0.3)
- P(X = 0.3)
- P(0.3 < X $\leq$ 1.3)
- P(0.2 $\leq$ X $\leq$ **or** 0.7 $\leq$ X $\leq$ 0.9)
- P(X is not in the interval (0.4, 0.7))

Answers

- $\sqcap = (0.7) \times (1) = 0.7$
- $\sqcap = 0$
  - The probability of being exactly on a point in the infinite sum will **always** be 0.
- $\sqcap = (0.7) \times (1) = 0.7$
  - Do not keep shading when there is no density curve, meaning it is a hard stop at $X = 1$
- $\sqcap = ((0.25 - 0.2) \times \frac{1}{0.25-0.2}) + ((0.91 - 0.7) \times \frac{1}{0.91-0.7}) = 0.25$
- $\sqcap = 0.4 + 0.3 = 0.7$

Normal Random Variable

$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$ $X \sim N(\mu\,\sigma)$

Density curve is also a "bell curve"

Empirical (68-95-99.7) Rule

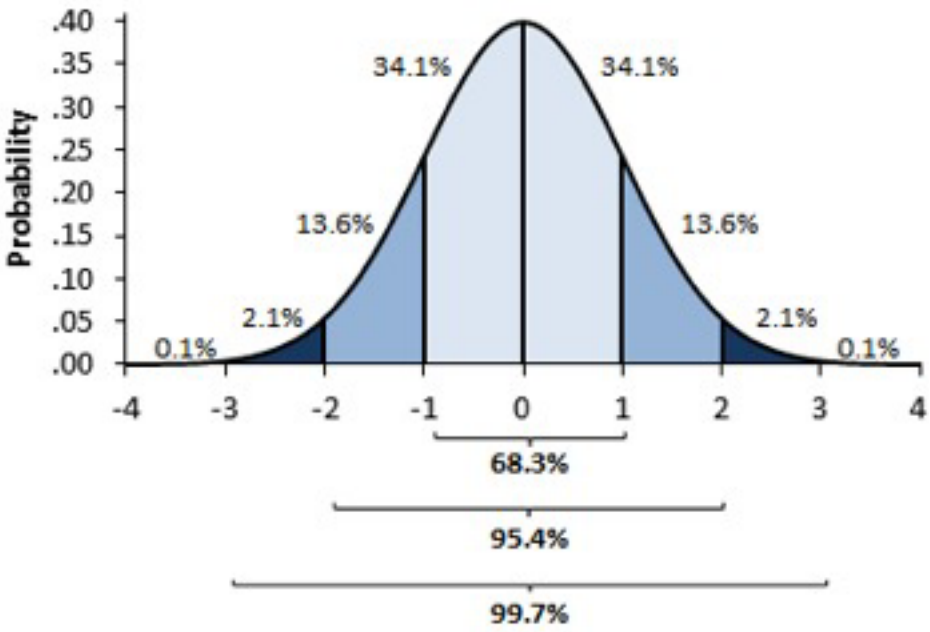


Figure 2: Bell Curve

$X \sim N(\mu\,\sigma)$

Standardization

It may be useful to standardize distributions to compare 2 variables with same density curve shape but different scales. For normal distributions $X \sim N(\mu, \sigma)$, we convert to Z-Scores $Z \sim N(0, 1)$

$$Z = \frac{x - \mu}{\sigma} = \frac{value\ -\ mean\, of\, distribution}{standard\, deviation}$$

$$P(Z \leq z) = P(X \leq x)$$

"Cumulative proportion"/"Cumulative probability"

Contents

Day 15

Z-Score Example

Two tests of "English" ability

- NAEP Reading Test
- SAT verbal

Suppose a student score scored 320 on NAEP & 650 SAT

Which test did he do better on?

$NAEP \sim N(288, 38)$

$SAT \sim N(500, 120)$

Convert to Z-Scores

NAEP:

$$Z = \frac{value - mean}{standard\,deviation} = \frac{320 - 288}{38} = 0.842$$

Student scored 0.842 standard deviation above average

SAT:

$$Z = \frac{value - mean}{standard\,deviation} = \frac{650 - 500}{120} = 1.25$$

Student scored 1.25 standard deviation above average

R-Code

```
pnorm(320, mean = 288, sd = 38)
[1] 0.8001355
```

Cumulative proportion of 0.800 (80%) which means 80^{th} percentile.

```
pnorm(650, mean = 500, sd = 120)
[1] 0.8943502
```

Cumulative proportion of 0.8943502 which means 89^{th} percentile.

Water bottle example

Questions

- Why does it continue to overfill
  - How much does it actually pour → average
- Why does Dr. Wynne have such terrible reaction speed?
  - Reaction speed → average
- Does the water fill at the same rate
  - Average rate for one pour
  - → average over several attempts

Expected value = μ = expected amount filled

$\bar{X}$ = average amount filled in a sample of pours "sample mean".

Variability: how variable are the individual values. (range)

- σ = Standard Deviation
- σ^2 = Variance
- S = Sample Standard Deviation
- S^2 = Sample Variance

Bias: Center: - on average, are we where we expected to be? (mean, median, mode)

Shape

Shape: where "average" is compared to "most likely"

- How "consistent" the values are given variability

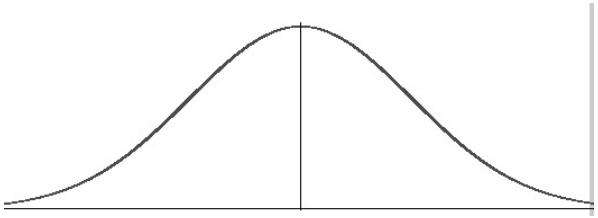


Figure 1: Unimodal Distribution

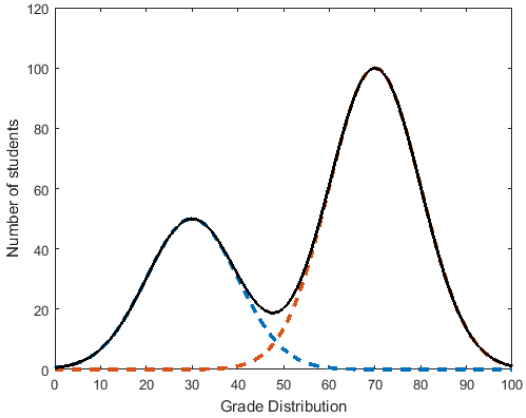


Figure 2: Bimodal Distribution

The median is resistant, the mean is subject to more change.

Figure 3: Left and Right Skewed Graphs

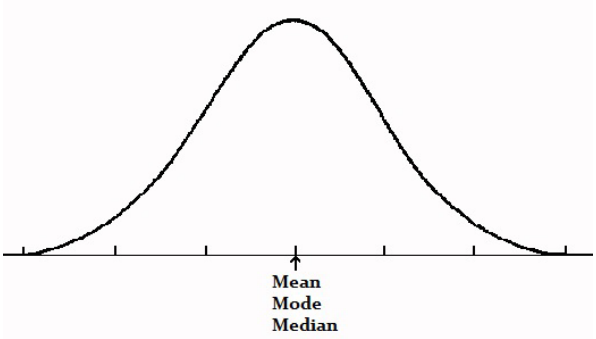


Figure 4: Symmetric Graph

Approximating a Density Curve: Histogram

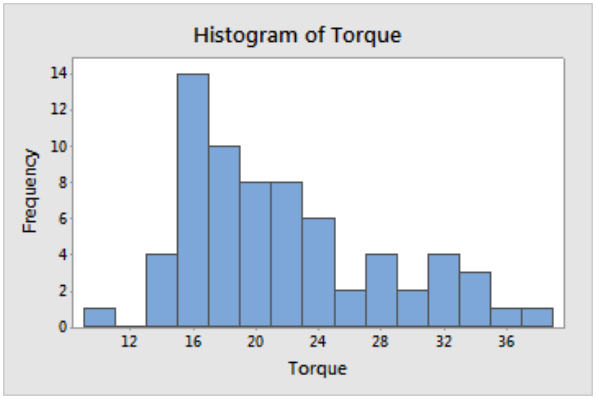


Figure 5: Histogram

- "bins": intervals on the x-axis
- Choice of bins is very important
  - Endpoints of bins
  - Center & Width
- Reimann Integral of an unknown density curve

Outliers

Points that doesn't fit with everything else

Attempting to determine outliers

- Plot your data & look for points that don't belong
- ↑ best way
- Investigate why they're different

Box-Plots

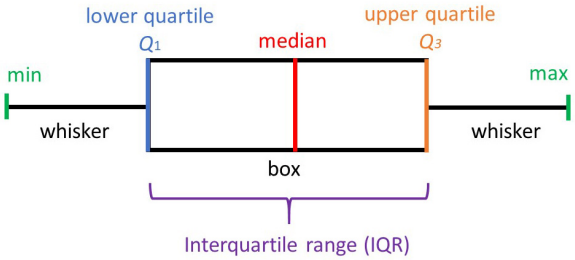


Figure 6: Box Plot

Rule of Thumb

- Step 1: Get five number summary (min, Q_1, medium(Q_2), Q_3, max)
- Step 2: Compute IQR = range middle 50% of data
  - IQR = Q_3 - Q_1
- Step 3: Compute "fences"
  - Lower fence: $Q_1 - K \times IQR$
  - Upper fence: $Q_3 + K \times IQR$

Anything outside the fences is an outlier.

By convention, $k = 1.5$

Example : Senator Ages

Five number summary:

- Min = 39
- Q_1 = 55.5
- Median = 63
- Q_3 = 69
- Max = 85

$IQR = 69 - 55.5 = 13.5$

Lower fence: $55.5 - (1.5)(13.5) = 35.25$ Upper fence: $69 + (1.5)(13.5) = 89.25$

In this data set we have no outliers because our data falls between the fences.

Numerical Variable Connection to Random Variables

Recall for random variable X

$E(A + Bx) = a + b \times E(x)$

$Var(A + Bx) = b^2 \times Var(x)$

$SD(A + Bx) = |b| \times sd(x)$

Recall for random variable X and Y

$E(Ax + By) = aE(x) + bE(y)$

$Var(Ax + By) = A^2 \times var(x) + B^2 var(y)$

$SD(Ax + By) = \sqrt{A^2 \times var(x) + B^2 \times var(y)}$

All of these rules hold for numerical variables too

Contents

Day 16

Error and Variability

Rounding variability: error due to precision of our machine/scale/etc.

1) Never measure exact, only to some tolerance
   - $\rightarrow$ typically rounding error $\sim (E, -E)$

Example: weight is 150 pounds - reality $\rightarrow 150 \pm U(-0.5, 0, 5)$

2) When making repeated measurements of something, there will be some natural variability, due to many small sources of error Usually (as long as errors are on the same scale), we can make measurement error of $\sim N(0, \sigma)$

3) Sampling error: error due to only having a sample from the population. Estimate a population mean μ based on a sample mean:

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} x_{\mathrm{i}}$$

i_

What is the distribution of $\bar{X}$ over all possible samples = sampling distribution of $\bar{x}$ which is the sampling mean

Consider simple random sample from a given population, the values $\mathrm{x}_1 \rightarrow \mathrm{x}_\mathrm{n}$ values of some numerical variables. We assume the x_i's are independent and identically distributed random variables.

With theoretical/population mean μ and standard deviation σ, then $\bar{X} = \frac{1}{n} \rightarrow (X_1 + X_2 + ... + X_\mathrm{n})$ is considered a linear combination.

$$E(\bar{X}) = E(\frac{1}{n} \times (X_1 + X_2 + ... + X_\mathrm{n})) = \frac{1}{n}(E(x_1) + E(x_2) + ... + x_\mathrm{n})$$

$$E(\bar{X}) = n\mu \times \frac{1}{n} = \mu$$

Important Facts

- The mean of the sampling distribution of $\bar{X}$ is equal to the population mean μ

$$(Var\bar{X}) = var(\frac{1}{n} \times (x_1 + x_2 + ... + x_\mathrm{n}))$$
$$= (\frac{1}{2})^2 \times var(x_1 + x_2 + ... + x_\mathrm{n})$$
$$= (\frac{1}{n})^2 \times var(x_1) + var(_2) + ... + var(x_\mathrm{n})$$

- The variance of sampling distribution of $\bar{X}$ is smaller than the population variance by a factor of n. The standard deviation is smaller by a factor of $\sqrt{n}$. Consider a normally distributed population. Theorem: any linear combination of normal random variables is also normally distributed.

$$\bar{X} \sim N(\mu, \frac{\sigma}{\sqrt{n}})$$

Central Limit Theorem

For a simple random sample (ASRS) of size `n` from a population with finite mean μ and finite standard deviation σ:

When `n` is "large enough", $\bar{x}$ is approximately $\sim N(\mu, \frac{\sigma}{\sqrt{n}})$

What does "large enough" depend on?

- How good the approximation needs to be (robust procedures-approximation just needs to be OK)
- Shape of population distribution
  - Higher skew requires larger `n`
  - Outliers in sample suggest larger `n` is needed

Consider a normally distributed population.

$\bar{X}$ is a linear combination of random variables $X_i \sim N(\mu, \sigma)$

So:

Example

You take a sample size of 64 from a population normally distributed with mean of 82 and standard deviation of 24.

a) Find the sampling distribution of the sample mean $\bar{X}$
b) Middle 95% of values of x are expected to be in what interval.
c) Middle 95% of sample means $\bar{X}$ are expected to be in what interval?

Answers

a) $\bar{X} \sim N(82, \frac{24}{\sqrt{64}}) \sim N(82, 3)$
b) (34, 130)
c) $E[\bar{x}] = \mu = 82$, $SD[\bar{x}] = \frac{\sigma}{\sqrt{n}} = 3$ $\mu + 2SD[\bar{x}] = 82 + 6 = 88$ $\mu - 2SD[\bar{x}] = 82 - 6 = 76$ The range between 76 and 88

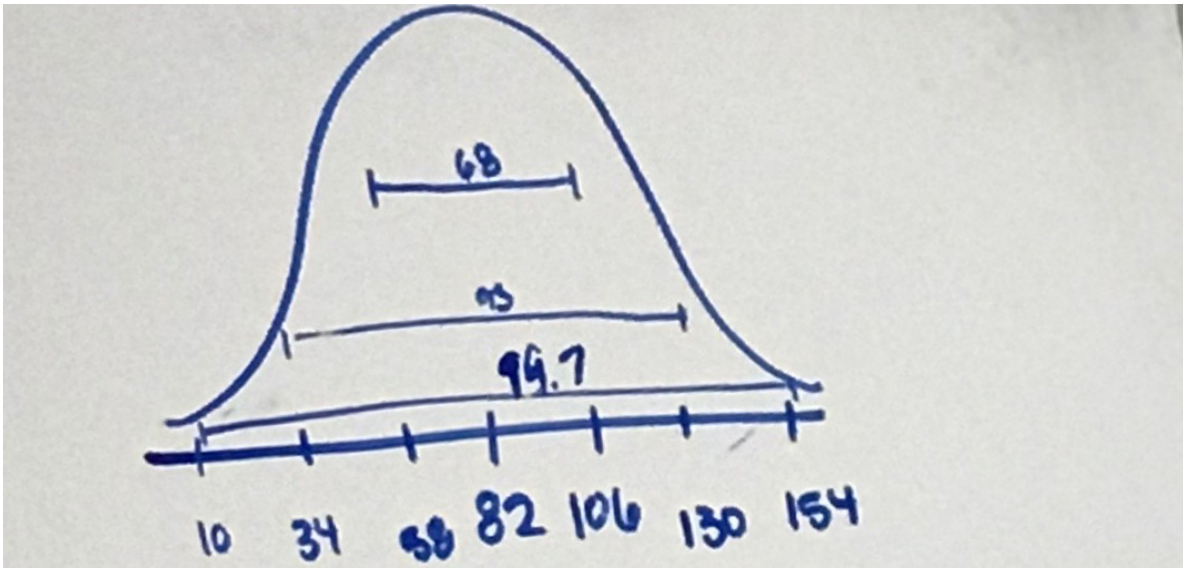


Figure 1: Curve