Introduction to Statistics

Danielle Maddix Robinson

Senior Applied Scientist
AWS AI Labs

July 24 - 25, 2023

Introduction to Statistics Overview

Instructor: Danielle Maddix Robinson
Workshop Assistant: Eden Luvishis

Main Topics:

- Descriptive statistics and sampling
- Probability and distributions
- Sampling distributions and central limit theorem
- Confidence intervals and hypothesis testing
- Correlation and Regression

Resources

Notes and other materials are posted:
`https://eluvishis.github.io/Stanford-ICME-Summer-2023-Stats/`

Course materials are based on past materials from:

- Prof. James Lambers
- Prof. Guenther Walther

Recommended textbook for further details:

- *The Elements of Statistical Learning: Data Mining, Inference and Prediction* by Hastie, Tibishrani, Friedman
- *An Introduction to Statistical Learning* by James, Witten, Hastie, Tibshirani
- *Probability* by Pitman

Descriptive Statistics

Descriptive Statistics summarize and display data so that it can be readily interpreted

Examples:

- Average, or mean, is a convenient way of describing a set of many numbers with a single number
- Variance is used to measure how far the set is from its mean
- Charts are useful for organizing and summarizing data

Inferential Statistics

Inferential statistics make claims about an entire (large) population based on a (relatively small) sample of data

Related topics:

- Confidence intervals
- Hypothesis testing
- Goodness-of-fit tests
- Correlation and regression

Statistics Example in Every Day Life

Suppose a pollster wanted to determine the percentage of all registered voters in California that would support a certain ballot measure

- Not be practical to question the entire population consisting of all of these voters, as there are millions of them
- Instead pollster questions a sample consisting of a reasonable number of these voters (e.g., 200 voters) and then use inferential statistics to make a conclusion about the voting preference of the entire population based on the data obtained from the sample

Distinction: Size of the population

Descriptive statistics:
Conclusions made about a relatively small population based on direct observations of every member of that population

Inferential statistics:
Conclusions made about a relatively large population based on descriptive statistics applied to a small sample from that population

Ethics in Statistics

In order to draw sound conclusions about a large population from a small sample:

- Sample of that population be representative of that population
- Otherwise sample is said to be biased

Example: 1936 Presidential Election

Poll of a sample of voters was conducted to determine whether the majority would vote for:

1. Democratic candidate: Franklin D. Roosevelt
   - Poll incorrectly predicts lost
   - In reality won
2. Republication candidate: Alf Landon
   - Poll incorrectly predicts won
   - In reality lost

Where Did The Polling Go Wrong?

Method of polling led to an unintentional bias

- Telephone directories used to obtain voter names
- In 1936, telephones existed primarily in more affluent households
- Those tended to vote Republican

Mean

Given a set of n numerical observations $\{x_1, x_2, \ldots, x_n\}$ of a population, the mean of the set is

$$\mu = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

When the observations are drawn from a sample, rather than an entire population, then the mean is denoted by $\bar{x}$:

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n}$$

The mean can be defined more concisely using sigma notation:

$$\mu = \frac{1}{n} \sum_{i=1}^{n} x_i$$

Median

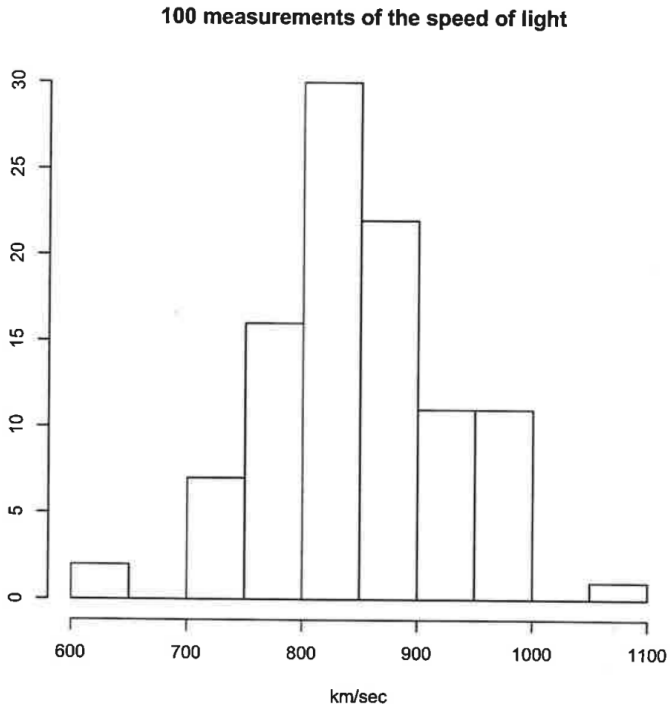
The median is the number that is larger than half the data and smaller than the other half, i.e., the 50th percentile

For a sorted set X of n observations, Median(X):

- $X[\frac{n+1}{2}]$, n odd (Middle Number)
- $\frac{X[\frac{n}{2}]+X[\frac{n}{2}+1]}{2}$, n even (Average of two middle values)

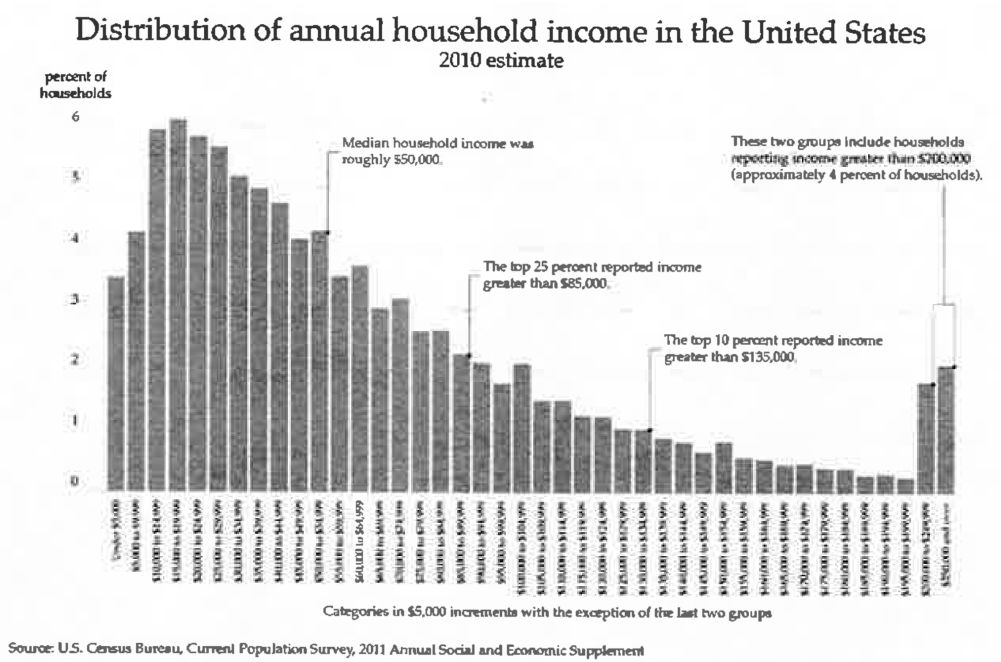
Mean vs. Median

Mean and median are the same when the histogram of the data is symmetric



Mean vs. Median

When histogram is skewed to the right, the mean can be much larger than the median



When histrogram is skewed, use median

Measures of Data Dispersion

- Provides more information about the dataset than just a numerical value, i.e., mean or median
- Values may be clustered closely around the mean or median, or they may be widely spread out
- Describes how far individual values deviate from the mean or median

Range

The range of a set of data observations is simply the difference between the largest and smallest values

It uses very little of the data, and is unduly influenced by outliers

Interquartile Range (IQR)

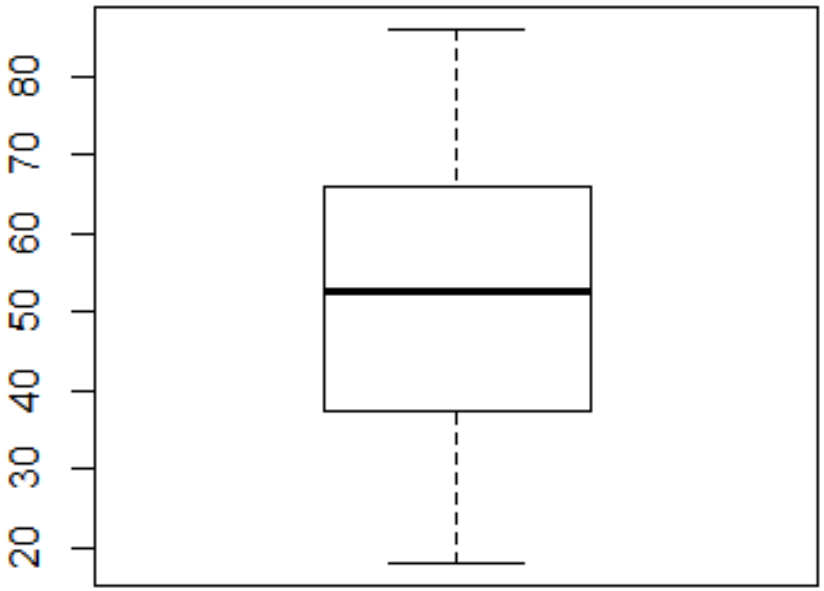
$$IQR = Q_3 - Q_1$$

- Q_1: median of the "lower half" of data, i.e., 25th percentile
- Q_3: median of the "upper half" of the data, i.e., 75th precentile

- Measures how spread out the center of data is
- Similar to median, is robust to outliers

Five-point Summary and Box Plot

Five-point summary: minimum, Q_1, median (Q_2), Q_3, and maximum

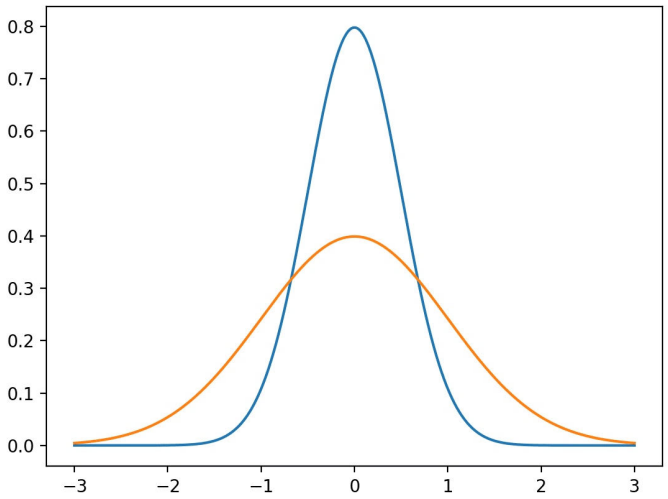


Less information than histogram but useful to compare datasets

Population Variance

The variance of a population is obtained from the deviation of each observation from the mean:

$$\sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2$$



Blue: small variance; orange: large variance

Sample Variance

The variance of a sample is slightly different:

$$s^2 = \frac{1}{N-1}\sum_{j=1}^{N}(x_j - \bar{x})^2$$

- Uses sample mean $\bar{x}$
- Division by $N-1$ instead of N compensates for sample variance when dividing by N to underestimate the population variance

Standard Deviation

For both a population and a sample, the standard deviation is the square root of the variance, i.e.,

$$\sigma = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(x_j - \mu)^2}$$

$$s = \sqrt{\frac{1}{N-1}\sum_{j=1}^{N}(x_j - \bar{x})^2}$$

An advantage is that the standard deviation is measured using the same units as the original data

Methods of Sampling

Once the determination is made that only a sample of a population of interest can be studied, how to obtain that sample is far from a trivial matter

It is essential that the sample not be biased; that is, the sample must be representative of the entire population, or any inferences made from the sample will not be reliable

To reduce the chance of bias, it is best to use random sampling, which means that every member of the population has a chance of being selected

We now discuss various approaches to random sampling

Simple Sampling

In simple sampling, each member of the population has an equal chance of selection

Typically, tables of random numbers are used to assist in such a selection process

For example, suppose all members of the population can be numbered. Then, the table of random numbers can be used to determine the numbers of members of the population who are to be included in the sample

Systematic Sampling

Simple sampling is susceptible to bias, if some aid such as a table of random numbers cannot be used

To avoid this bias, one can use systematic sampling, which consists of selecting every kth member of the population

If the population has N members and a sample of size n is desired, then one should choose $k \approx N/n$

Cluster Sampling

In cluster sampling, the population is divided into groups, called clusters, and then random sampling is applied to the clusters

That is, entire clusters are chosen to obtain the sample

This is effective if each cluster is representative of the entire population

Stratified Sampling

In stratified sampling, the population is divided into mutually exclusive groups, called strata, and then random sampling is performed within each stratus

This approach can be used to ensure that each stratus is treated equally within the sample

For example, suppose that for a national poll, it was desired to have a sample in which each state was represented equally

Then, the strata would be the states, and a sample could be obtained from the populations of each state

Sampling Pitfalls

Sampling must be performed with care, so that any inferences made about the population from the sample have at least some validity

Sampling Errors

A descriptive statistic computed from a sample is only an estimate of the corresponding statistic for the population, which in most cases cannot be obtained

However, it is possible to estimate the error in the sample statistic, called the sampling error; we will learn how to do so later, using confidence intervals

As we will see then, choosing a larger sample reduces the sampling error. It can be made arbitrarily small by choosing a sample close to the size of the entire population, but usually this is not practical

Poor Sampling Technique

Even if a very large sample is chosen, conclusions made about the sample do not apply to the population if the sample is biased

On the other hand, if a sample is truly representative of the population, then it does not need to be large to be reliable

It is also important to avoid making unrealistic assumptions about the sample

1948 Presidential Election

In a poll conducted during the 1948 presidential election, voters in the sample were classified as supporting Harry Truman, supporting Thomas Dewey, or undecided

The polling organization made the assumption that undecided voters should be distributed among the two candidates in the same way that the decided voters were, which led to a conclusion that Dewey would win

However, the undecided voters were actually more in favor of Truman, thus leading to his victory

What is Probability?

In 2015, there were about 4 million babies born in the US, and 48.8% of newborns were girl

$$\text{P(newborn is girl)} = 48.8\%$$

Probability of an event: proportion of times this event occurs out all possible outcomes of an experiment

In WWII, John Kerrich tossed a coin 10,000x and observed 5067 heads

$$\text{P(head)} = \frac{5067}{10000} = 50.67\%$$

1. Complement Rule

Probabilities are between 0 and 1

$$P(\text{A does not occur}) = 1 - P(A)$$

Write A for event, i.e., A = 'newborn is girl'

$$P(\text{newborn is girl}) = 48.8\%$$

$$P(\text{newborn is boy}) = 51.2\%$$

2. Equally Likely Outcomes

If there are n possible outcomes and they are equally likely, then

$$P(A) = \frac{\text{number of outcomes in A}}{n}$$

Rolling a dice

- Since each of its 6 faces is equally likely, each has probability 1/6

3. Addition Rule

If A and B are mutually exclusive, then

$$P(A \text{ or } B) = P(A) + P(B)$$

Events A and B are mutually exclusive if they cannot occur at the same time

Example: Roll a die twice

- A = 1 on 1st roll
- B = 6 on 1st roll
- C = 1 on 2nd roll

$\rightarrow$ A and B are mutually exclusive, but A and C are not

4. Multiplication Rule

If A and B are independent, then

$$P(\text{A and B}) = P(A)P(B)$$

Events A and B are independent if knowing that one occurs does not change the probability that the other occurs

Example: Roll a die twice

- A = 1 on 1st roll
- B = 6 on 1st roll
- C = 1 on 2nd roll

$\rightarrow$ B and C are independent, but A and B are not

4 Rules Example

Roll a dice 3x. What is P(at least one 6)?

$$\text{P(at least one 6)} = \text{P((6 roll 1) or (6 roll 2) or (6 roll 3))}$$

Events are not mutually exclusive so cannot use addition rule

Use complement rule and multiplication rule

$$\begin{aligned}
\text{P(at least one 6)} &= 1 - \text{P(no 6 in 3 rolls)} \\
&= 1 - \text{P((no 6 roll 1) and (no 6 roll 2) and (no 6 roll 3))} \\
&= 1 - \text{P(no 6 roll 1)P(no 6 roll 2)P(no 6 roll 3)} \\
&= 1 - \frac{5}{6}\frac{5}{6}\frac{5}{6} \\
&= 42.1\%
\end{aligned}$$

Conditional Probability

Conditional probability of B given A:

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

General multiplication rule: $\mathrm{P(A \text{ and } B)} = \mathrm{P(A)P}(B|A)$

If A and B are independent: then

$$P(A \text{ and } B) = P(A)P(B)$$

$$\mathrm{P}(B|A) = \mathrm{P(B)}$$

Conditional Probability Example

Spam email has a higher chance to contain the word 'money' than ham (not spam) email:

$$P(\text{money} \mid \text{spam}) = 8\%, \quad P(\text{money} \mid \text{ham}) = 1\%$$

$$P(\text{spam}) = 20\%$$

What is the probability that 'money' appears in an email?
Idea is to artificially introduce even spam/ham

$$\begin{aligned} P(\text{money appears}) &= P(\text{money and spam}) + P(\text{money and not spam}) \\ &= P(\text{money} \mid \text{spam})P(\text{spam}) + P(\text{money} \mid \text{ham})P(\text{ham}) \\ &= 0.08 * 0.2 + 0.01 * 0.8 \\ &= 2.4\% \end{aligned}$$

Bayes' Rule

Given two events A and B, Bayes' Rule is a relates the conditional probabilities $P(A|B)$ and $P(B|A)$:

It states that

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

By the multiplication rule, the numerator on the right-hand side is simply $P(A \cap B)$, and the denominator becomes $P(A \cap B) + P(A \cap B')$

Bayes' Rule cont'd

Because B and B' are mutually exclusive, but also exhaustive (meaning $B \cup B'$ is equal to the entire sample space), this expression becomes $P((A \cap B) \cup (A \cap B')) = P(A)$

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

which can be rearranged to again obtain the multiplication rule

Alternative Form of Bayes' Rule

If we keep the original numerator in but use the simplified denominator, we obtain another commonly used form of Bayes' Rule:

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

This form is very useful for computing one conditional probability from another that may be easier to obtain

Spam Example

In prior example, we saw from data that $P(\text{money} \mid \text{spam}) = 8\%$

We need $P(\text{spam} \mid \text{money})$ to build a spam filter

Use Bayes' Rule!

$$\begin{aligned} P(\text{spam} \mid \text{money}) &= \frac{P(\text{money} \mid \text{spam})P(\text{spam})}{P(\text{money})} \\ &= \frac{P(\text{money} \mid \text{spam})P(\text{spam})}{P(\text{money} \mid \text{spam})P(\text{spam}) + P(\text{money} \mid \text{ham})P(\text{ham})} \\ &= \frac{0.08 * 0.2}{0.024} \\ &= 67\% \end{aligned}$$

Bayesian Analysis

Spam filter classifies email as spam via Bayesian analysis

- Before examining the email, there is a prior probability $P(B)$ of 20% that it is spam
- After examining the email for certain keywords, e.g., money, the filter updates this prior using Bayes' Rule to arrive the posterior $P(B|A)$ of 67% that the email is spam

Insurance Example

Suppose an insurance company classifies people as accident-prone or not accident-prone

Probability of an accident-prone person actually having an accident within the next year is 0.4, whereas the probability of a non-accident-prone person having an accident within the next year is 0.2

If 30% of people are accident-prone, then what is the probability that someone who does have an accident within the next year actually is accident-prone?

Applying Bayes' Rule

Let A be the event that the person has an accident within the next year, and let B be the event that the person is accident-prone

From the given information, we have

$$P(A|B) = 0.4, \quad P(A|B') = 0.2, \quad P(B) = 0.3$$

From these probabilities, we conclude that

$$P(A) = P(A|B)P(B) + P(A|B')P(B') = (0.4)(0.3) + (0.2)(0.7) = 0.26$$

Using Bayes' Rule, we conclude that the probability of someone who has an accident being accident-prone is

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{(0.4)(0.3)}{0.26} = 0.4615$$

Introduction

Now that we know how to compute probabilities of events, we study the behavior of the probability across all possible outcomes of an experiment, i.e., the distribution of the probability across the sample space

Our understanding of the probability distribution allows us to make inferences from the data from which the distribution arises

Random Variables

A random variable, X, is an outcome of an experiment that has a numerical value

The value itself is usually denoted by the lower-case version of the letter used to denote the variable itself, i.e., a random variable X takes on numerical values x

Random variables can either be discrete or continuous

Discrete Probability Distributions

A discrete probability distribution is a listing of all possible values of a discrete random variable, along with the probability of each value being assumed by the variable

Example

Let X be a discrete random variable whose outcomes correspond to the place one finishes in a race:

- first
- second
- third, etc.

If there are 10 runners in the race, then X can assume as a value any positive integer between 1 and 10

The Distribution

Example probability distribution:

x	$P(X = x)$
1	0.1
2	0.15
3	0.23
4	0.18
5	0.15
6	0.1
7	0.04
8	0.02
9	0.02
10	0.01

$P(X = x)$ denotes the probability that X assumes the value x

Rules for Discrete Distributions

1. Each outcome must be mutually exclusive of the others; i.e., X cannot assume two values simultaneously as the result of an experiment
2. For each outcome x, we must have $0 \leq P(X = x) \leq 1$
3. If the distribution has n possible outcomes $x_1, x_2, \ldots, x_n$, then we must have

$$\sum_{i=1}^{n} P(X = x_i) = 1$$

Expected Value (Mean)

"Most likely", or expected value, that the variable assumes of a probability distribution

Weighted mean of the outcomes, where the probabilities serve as the weights

Mean, or expected value, of the discrete random variable X:

$$E[X] = \mu = \sum_{i=1}^{n} x_i P(X = x_i)$$

Example

Consider a raffle, in which each ticket costs \$5

There is one grand prize of \$100, two first prizes of \$50 each, and four second prizes of \$25 each

If 200 tickets are sold, then the probability of winning the grand prize is $1/200 = 0.005$, while the probabilities of winning first prize and second prize are $2/200 = 0.01$ and $4/200 = 0.02$

Expected amount of winnings:

$$E[X] = 100(0.005) + 50(0.01) + 25(0.02) + 0(0.965) = 1.5$$

Interpretation

A ticket holder can expect to win, on average, \$1.50

However, we must account for the cost of the ticket, which applies to all participants; therefore, the expected net winnings is −\$3.50

Since the expected amount is negative, the raffle is not fair to the ticket holders

If the expected value was zero, then the raffle is a "fair game"

Variance

Using the mean of X, we can then characterize the dispersion of the outcomes by defining the variance of X as follows:

$$\sigma^2 = \sum_{i=1}^{n} (x_i - \mu)^2 P(X = x_i)$$

An equivalent formula, in terms of expected values, is

$$\begin{aligned} \sigma^2 &= E[X^2] - E[X]^2 \\ &= E[X^2] - \mu^2 \end{aligned}$$

Note that in the first term, the values of X are squared, and then they are multiplied by the probabilities and summed, whereas in the second term, the expected value is computed first, and then squared

Discrete Uniform Distribution

Uniform distribution $X \sim \mathcal{U}\{a, b\}$ is the probability distribution for a random variable X with domain $\{a, a+1, \ldots, b\}$

Each value in the domain of X is equally likely to be observed

Probability mass function:

$$P(X = k) = \frac{1}{n}, \quad n = b - a + 1, \quad k \in \{a, a+1, \ldots, b\}$$

Mean and Variance

Using the above definitions of the mean and variance of a discrete random variable, it can be shown that

$$E[X] = \frac{a+b}{2}, \quad Var[X] = \frac{(b-a+1)^2 - 1}{12}$$

Hint: Use sum identities

$$\sum_{i=0}^{n-1} i = \frac{n(n-1)}{2}, \quad \sum_{i=0}^{n-1} i^2 = \frac{n(n-1)(2n-1)}{6}$$

$$E[X] = \sum_{i=0}^{n-1}(a+i)P(X=a+i) = \frac{1}{n}\left(na + \sum_{i=0}^{n-1} i\right)$$

Binomial Experiments

Suppose that an experiment is performed n times, and it can have only two outcomes, that are classified as "success" and "failure"

Each of these individual experiments is a Bernoulli trial

- Each trial is independent of the others
- Probability of success: p, where $0 < p < 1$
- Probability of failure: $q = 1 - p$

Binomial Distribution

Binomial distribution $X \sim B(n, p)$ is the probability distribution for the discrete random variable X

- n: Bernoulli trials
- p: probability of success for each trial
- k: value of X, i.e., number of successes

Given a value for k, $0 \leq k \leq n$, what is $P(X = k)$?

Since the trials are independent, the probability of success (or failure) in consecutive trials is the product of the probabilities of the outcomes of each trial

Probability of k successes, followed by $n - k$ failures:

$$p^k(1-p)^{n-k}$$

Examples

- Testing for defective parts
  - n: number of parts to be checked
  - p: probability that a part is not defective
  - k: number of parts that are not defective
- Observing the number of correct responses on exam
  - n: number of questions
  - p: probability of getting the correct answer on a question
  - k: number of correct responses
- Counting number of households with an internet connection
  - n: number of households
  - p: probability of a household having an internet connection
  - k: number of households that have an internet connection

Probability Mass Function

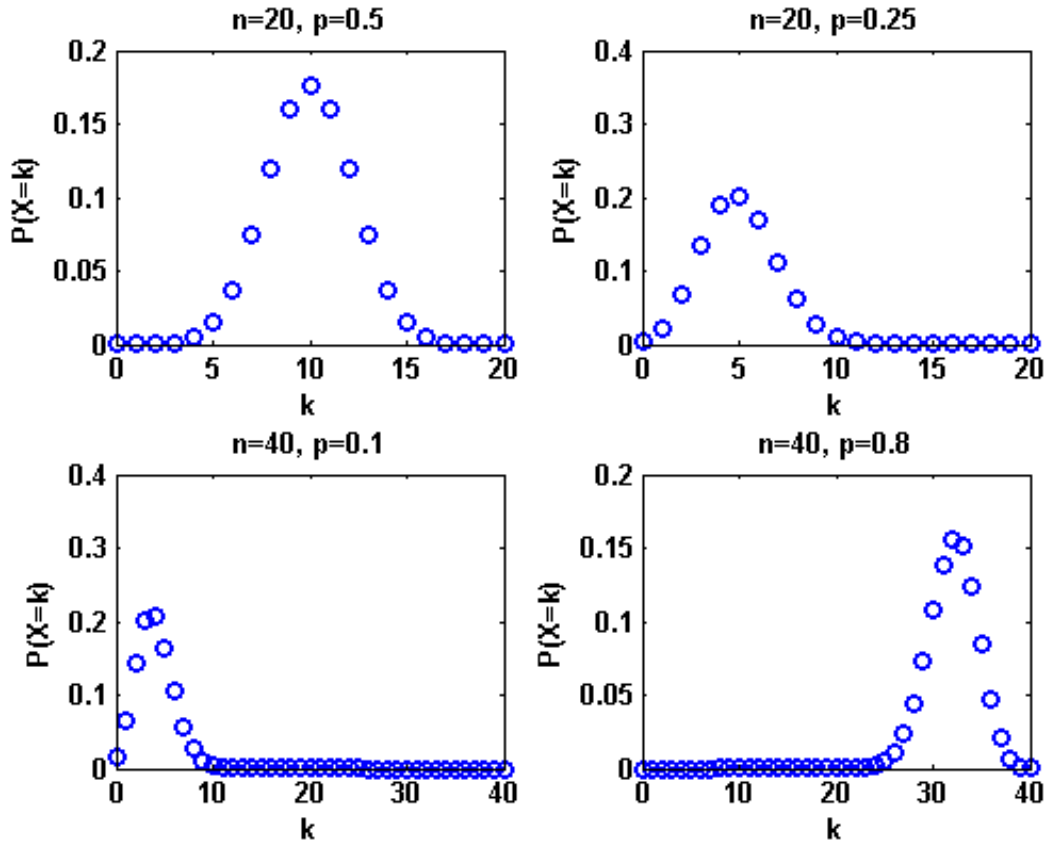
To determine the probability that any k of the n trials are successful, we have to consider all possible ways to choose k trials out of the n to be successful

We conclude that the probability mass function for the binomial distribution is

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k} = \frac{n!}{k!(n-k)!} p^k (1-p)^{n-k}$$

By the Binomial Theorem, it can be verified that the sum of all of these probabilities is 1

Examples



The binomial distribution, for various values of n and p

Behavior of the Binomial Distribution

- Symmetric if $p = 0.5$, in which case the probability mass function simplifies to $P(X = k) = \binom{n}{k} 2^{-n}$
- Skews left if $p < 0.5$ since there is a greater probability of more failures
- Skews right if $p > 0.5$ since there is a greater probability of more successes

Mean and Variance

Using the definition of expected value, linearity of expectation and sum of variances of independent trials:

$$E[X] = np, \quad Var[X] = npq$$

Hint: Write X as the sum of n Bernoulli random variables Y_i, where $E[Y_i] = p$, $Var[Y_i] = pq$

$$E[X] = E(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} E(Y_i) = \sum_{i=1}^{n} p = np$$

$$Var[X] = \underbrace{Var(\sum_{i=1}^{n} Y_i) = \sum_{i=1}^{n} Var(Y_i)}_{\text{independent trial assumption}} = \sum_{i=1}^{n} pq = npq$$

Continuous Probability Distribution

Continuous random variable is a random variable X whose domain is an interval $D = [a, b]$, which is a subset of the real numbers $\mathbb{R}$

Continuous probability distribution is a function $f : D \to [0, 1]$ whose value at $x \in D$ is the probability $P(X = x)$

Probability density function (pdf) of X is $f(x)$

Analogous with the requirement that the probability mass function in the discrete case must sum to one, the integral of the pdf must satisfy

$$\int_a^b f(x)\, dx = 1$$

Mean and Variance

The mean, or expected value, of a continuous random variable X:

$$E[X] = \int_a^b xf(x)\,dx$$

Then, we can define the variance in the same way as for a discrete random variable:

$$\text{Var}[X] = E[X^2] - E[X]^2$$

Continuous Uniform Distribution

Continuous uniform distribution $X \sim \mathcal{U}(a, b)$ is the probability distribution for a random variable X with domain $[a, b]$

All subintervals of $[a, b]$ of the same width are equally likely to be observed

Probability density function for this distribution:

$$f(x) = \frac{1}{b-a}, \quad x \in [a, b]$$

Mean, variance

$$E[X] = \frac{a+b}{2}, \quad \sigma^2 = \frac{(b-a)^2}{12}$$

Normal Distribution

Normal distribution $X \sim \mathcal{N}(\mu, \sigma^2)$ is a probability distribution that is followed by continuous random variables, that can assume any real value within some interval

A normal distribution has two parameters:

1. Mean μ
2. Variance σ^2

Its mean, median and mode are all the same, and equal to μ

Characteristics

The distribution is "bell-shaped" and is symmetric around the mean

Area under the entire bell-shaped normal distribution curve must be equal to 1

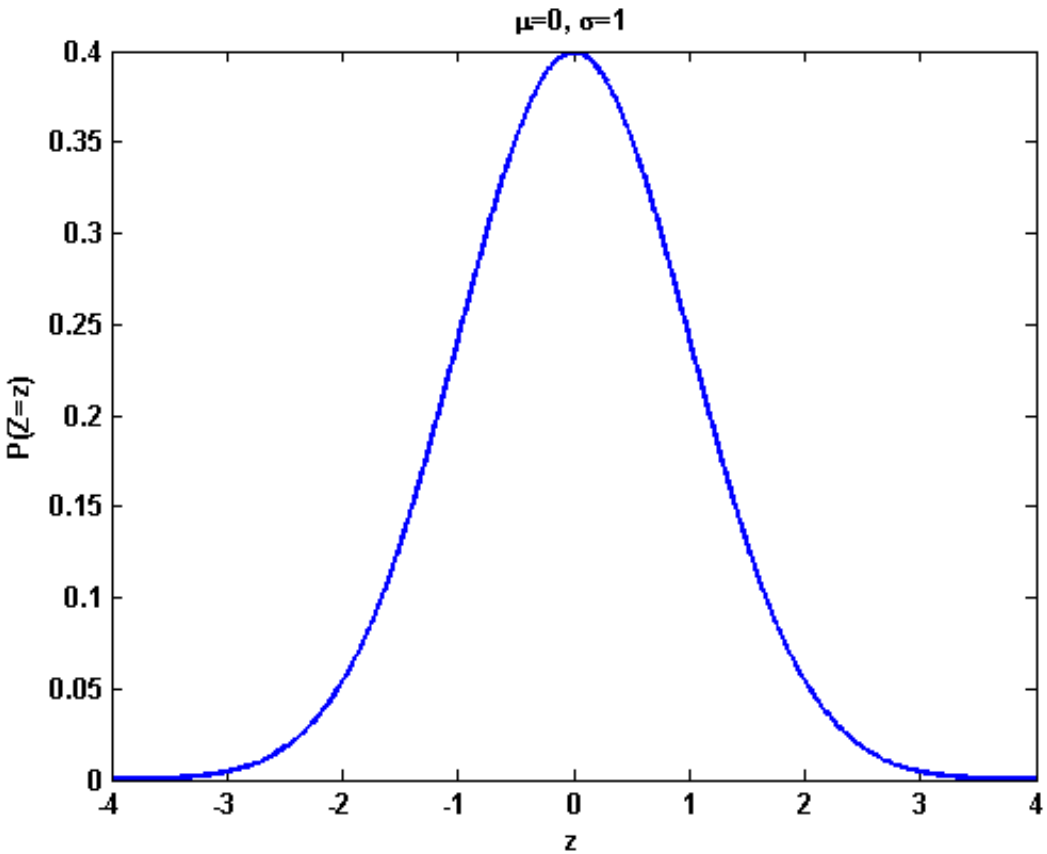
The probability is always strictly positive; it can never be zero

Probability approaches 0 for values that are far from the mean

Probability density function:

$$P(X = x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/(2\sigma^2)}$$

The Standard Normal Distribution



The standard normal distribution, with mean 0 and standard deviation 1

Calculating Probabilities

We wish to compute $P(X \leq x_0)$, i.e., the area of the region bounded by the normal distribution curve, the x-axis, and the vertical line $x = x_0$

$$P(X \leq x_0) = \int_{-\infty}^{x_0} P(X = x)\, dx = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{x_0} e^{-(x-\mu)^2/(2\sigma^2)}\, dx$$

Integral cannot be evaluated using analytical techniques from calculus

It must instead be evaluated numerically, which is cumbersome

How to Compute: Tables and z-scores

Tables use the standard normal distribution $\mathcal{N}(0, 1)$

Conversion to the standard distribution must be performed first using the z-score:

$$z = \frac{x - \mu}{\sigma}$$

If x is a value of the normal distribution $\mathcal{N}(\mu, \sigma^2)$, then z is the corresponding value in $\mathcal{N}(0, 1)$

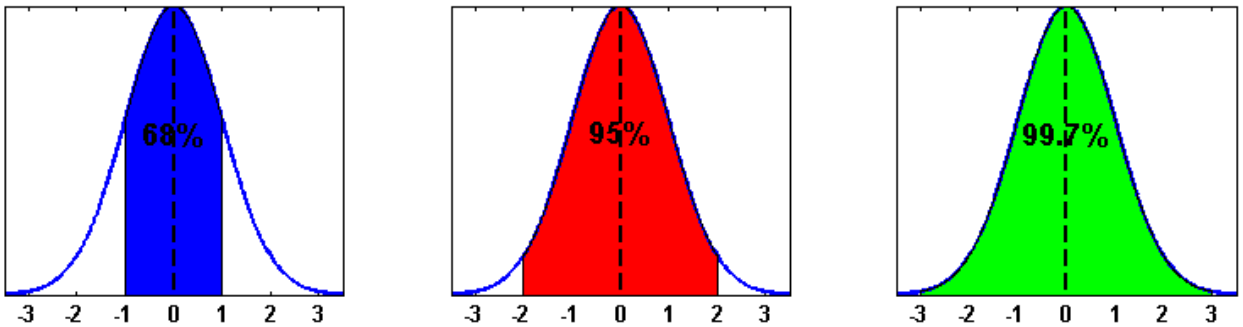
- z is the number of standard deviations between x and μ

Examples Using Symmetry

- $P(X \leq x_0)$: Obtain $P(Z \leq z_0)$ from a table, `pnorm` in R, `scipy.stats.norm.cdf` in Python (z_0 is the z-score for x_0)
- $P(X > x_0) = 1 - P(X \leq x_0)$ since events $X > x_0$ and $X \leq x_0$ are complementary
- $P(X \leq \mu - x_0) = P(X > \mu + x_0) = 1 - P(X \leq \mu + x_0)$ by symmetry
- $P(X > \mu - x_0) = P(X \leq \mu + x_0)$ by symmetry
- $P(x_1 \leq X \leq x_2) = P(X \leq x_2) - P(X \leq x_1)$

Empirical Rule

Empirical rule states that if the distribution of a set of observations is "bell-shaped", meaning that the distribution is symmetric around the mean and decreases toward zero away from the mean, then approximately 68, 95, and 99.7 % of the observations fall within 1, 2, and 3 standard deviations of the mean, respectively.



Empirical Rule for Normal Distribution

Empirical rule can be used to estimate normal distribution probabilities

While it is approximately true for any bell-shaped, symmetric distribution, it is exact for any normal distribution

The rule is derived from the behavior of the normal distribution

In terms of probabilities, the empirical rule states that

$$\begin{aligned} P(-1 \leq Z \leq 1) &\approx 0.68 \\ P(-2 \leq Z \leq 2) &\approx 0.95 \\ P(-3 \leq Z \leq 3) &\approx 0.997 \end{aligned}$$

Chebyshev's Theorem

Another rule of thumb, that applies even to distributions that are not bell-shaped or symmetric, is Chebyshev's Theorem, which states that if $k > 1$, then at least

$$\left(1 - \frac{1}{k^2}\right) 100\%$$

of the observations fall within k standard deviations of the mean

Other Probability Distributions

- Hypergeometric distribution: for sampling without replacement
- Exponential distribution: for studying time between events in a Poisson process
- Chi-square distribution: for goodness-of-fit, independence tests
- F-distribution: for analysis of variance
- and others...

Sampling Distributions

Sampling distribution: set of outcomes obtained from samples

Suppose we want to measure some quantifiable characteristic of a population, e.g., average height, or the percentage of the population that votes Republican

A sample of the population can be taken, and then the characteristic of the sample can be computed from information obtained from each member of the sample

Suppose many samples are taken with each sample being the same size

The values that are computed from these samples form a set of outcomes, where the experiment is the computation of the desired characteristic of the sample

Sampling Distribution of the Mean

Sampling distributions apply to a number of different statistics

Most commonly used is the mean

Sampling distribution of the mean is the pattern of means that is obtained from computing the sample means from all possible samples of the population

Example

Sampling distribution of the mean for an example of rolling a six-sided die

Each of the six numbers has an equal likelihood of appearing face up

These values follow a discrete uniform probability distribution that assigns the same probability to each discrete event

Example cont'd

The mean:

$$\mu = \frac{a+b}{2}$$

The variance:

$$\sigma^2 = \frac{1}{12}[(b-a+1)^2 - 1]$$

where a and b are the minimum and maximum values of the distribution

For a six-sided die, $a = 1$ and $b = 6$, $\mu = 3.5$ and $\sigma^2 = 35/12$

Example, cont'd

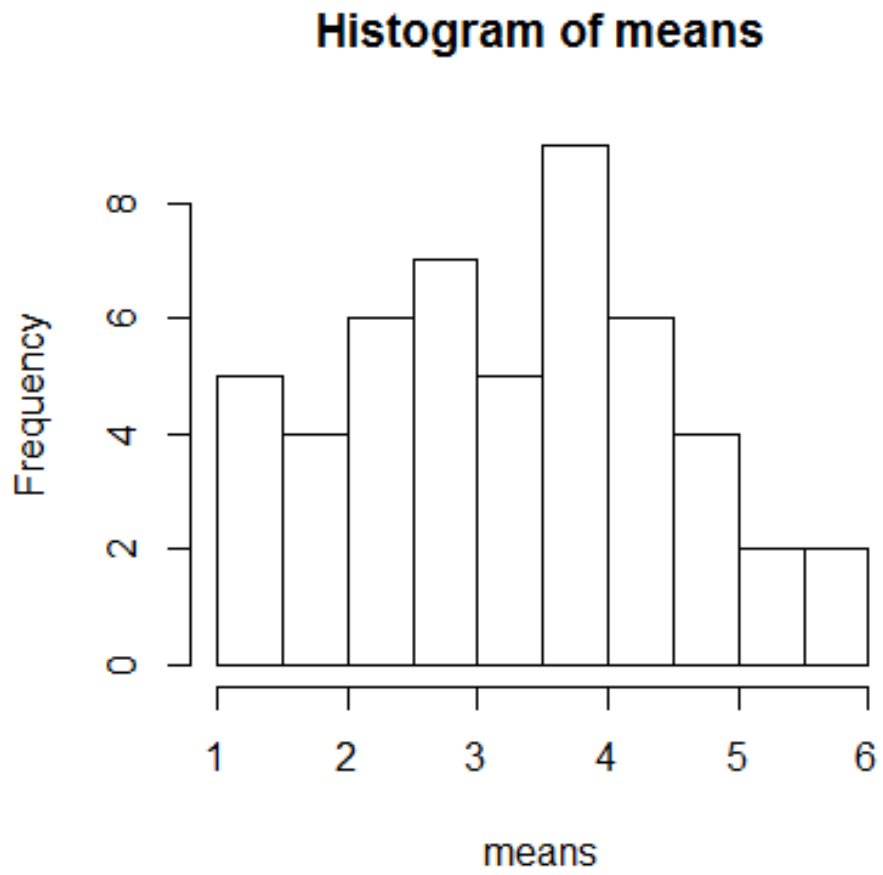
Suppose we roll the die n times, where n is the size of our sample, and compute the sample mean $\bar{x}$

We repeat this process m times, gathering m samples, each of size n

The m sample means $\{\bar{x}_1, \ldots, \bar{x}_m\}$ form a sampling distribution of the mean

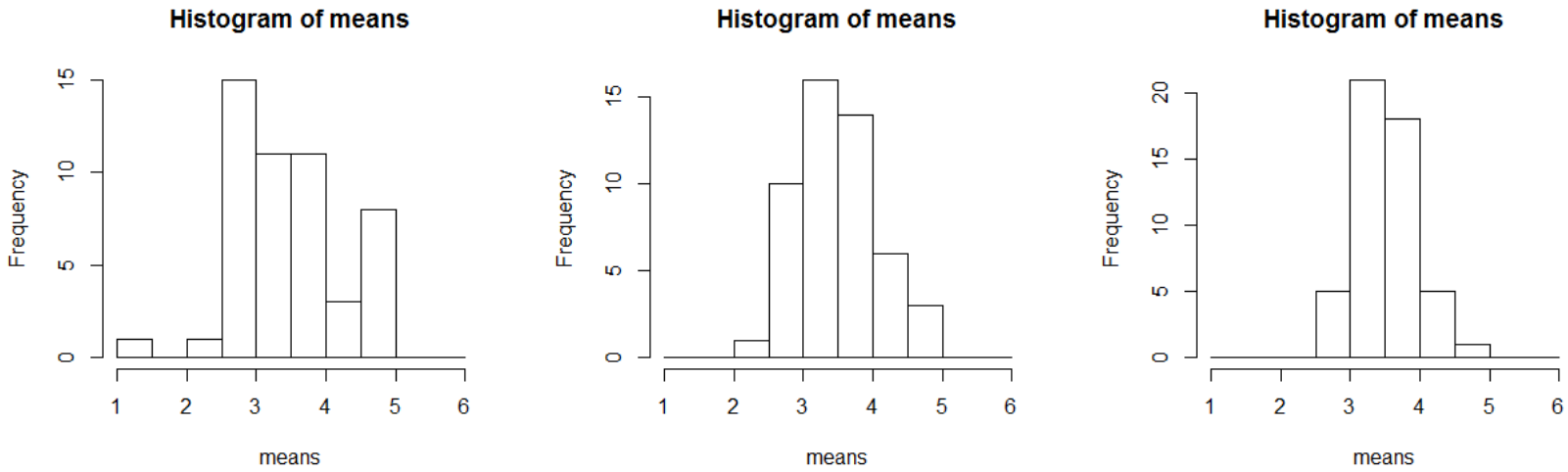
Sampling Distribution of the Mean: $n = 2$

Small sample of size $n = 2$, and compute $m = 50$ samples
The means are well-distributed across the interval from 1 to 6



Increasing the Sample Size n

Increase n (keeping m fixed) and see what happens to the distribution



For large n (*right*), distribution looks like normal distribution, with its mean roughly that of the original uniform distribution ≈ 3.5

Central Limit Theorem (CLT)

The behavior in the preceding example is no coincidence

Illustration of the Central Limit Theorem

This theorem states that as the sample size n increases, the sample means tend to converge to a normal distribution around the true population mean, i.e., $\mu_{\bar{x}} = \mu$

Holds regardless of the distribution of the population from which the sample is taken

Standard Error of the Mean

Central Limit Theorem also states that as the sample size n increases, the standard deviation of the sample means, $\sigma_{\bar{x}}$, converges to

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$$

where σ is the standard deviation of the population

Standard deviation of the sample means is called the standard error of the mean

In summary, for sufficiently large n

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

Example

In the case of the roll of a six-sided die, with a sample size of $n = 20$, the standard error is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{\sqrt{35/12}}{\sqrt{20}} = 0.382$$

Probability that the sample mean is greater than 4: $P(\bar{X} > 4)$
Compute the *z*-score for 4:

$$\frac{4 - \mu}{\sigma_{\bar{x}}} = \frac{4 - 3.5}{0.382} = 1.309$$

Then

$$P(\bar{X} > 4) = 1 - P(\bar{X} \leq 4) = 1 - P(Z \leq 1.309) = 1 - 0.9047 = 0.0953$$

Less than 10% chance that the sample mean is greater than 4

Sampling Distribution of the Sum

Suppose that instead of taking the mean of the observations in each sample, we take the sum

If the population mean and standard deviation are μ and σ, then as n increases, the sampling distribution of the sum converges to:

$$\mathcal{N}(n\mu, \sigma^2 n)$$

Mean and standard deviation of the sampling distribution of the mean are simply multiplied by n

Approximating the Binomial Distribution

Normal distribution can be used to approximate the binomial distribution

Sum of n random, independent, identically distributed (i.i.d.)

Bernoulli random variables with mean $\mu = p$, variance $\sigma^2 = p(1-p)$

By Central Limit Theorem,

$$Y \sim \mathcal{N}(np, np(1-p))$$

As long as the number of trials n and the probability of success p satisfy

$$np \geq 5, \quad n(1-p) \geq 5$$

Un-discretization

For computing probabilities, it is best to use the midpoints of the discrete values of the number of successes

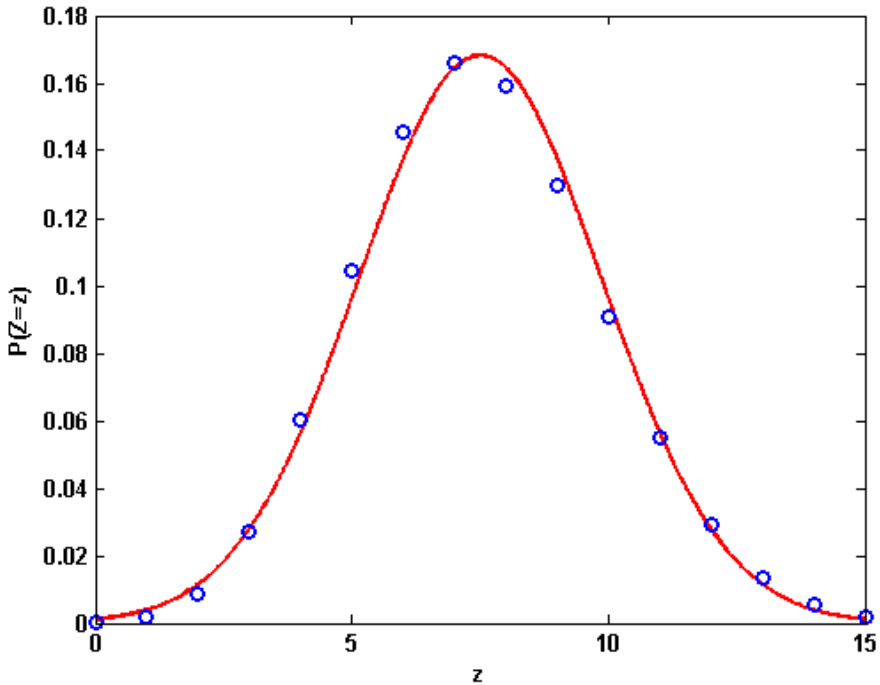
To approximate $P(X \leq 5)$, where X is a discrete random variable with a binomial distribution

Use continuous random variable $Y \sim \mathcal{N}(np, np(1-p))$

Compute $P(Y \leq 4.5)$ rather than $P(Y \leq 5)$

Due to the change from a discrete random variable to a continuous random variable

Example



Approximation of the binomial distribution with $n = 30$ and $p = 0.25$ (blue circles) by $\mathcal{N}(np, np(1-p))$ (red curve)

Introduction

Now that we have learned about sampling and sampling distributions, we are ready to learn how to use inferential statistics to make conclusions about populations based on information obtained from samples

A key component of inferential statistics is to quantify the uncertainty that is inherent in using only a sample

An example is polling: a statement of a poll result is accompanied by an indication of the sampling error

Confidence Intervals for Means

Suppose that we want to know the population mean and only have a sample mean

We can construct a confidence interval that is centered at the sample mean and can provide an indication of the population mean

Large Samples

We first consider the case where the sample size n is sufficiently large, i.e., $n \geq 30$

By the Central Limit Theorem, the sample means are approximately normally distributed, even if the population is not

Estimators

The sample mean is an example of a point estimate

- Single value that describes population
- Easy to compute, but hard to validate

To assess the validity of a sample mean, we rely on an interval estimate

- Range of values that describes the population
- Confidence interval: particular interval estimate we will use

Confidence Levels

The first step in constructing a confidence interval is choosing a confidence level, which is the probability that the interval estimate will include the population parameter (in this case, the population mean)

For example, for a 90% confidence interval, the confidence level is 0.9

Subtracting this value from 1 yields the significance level α

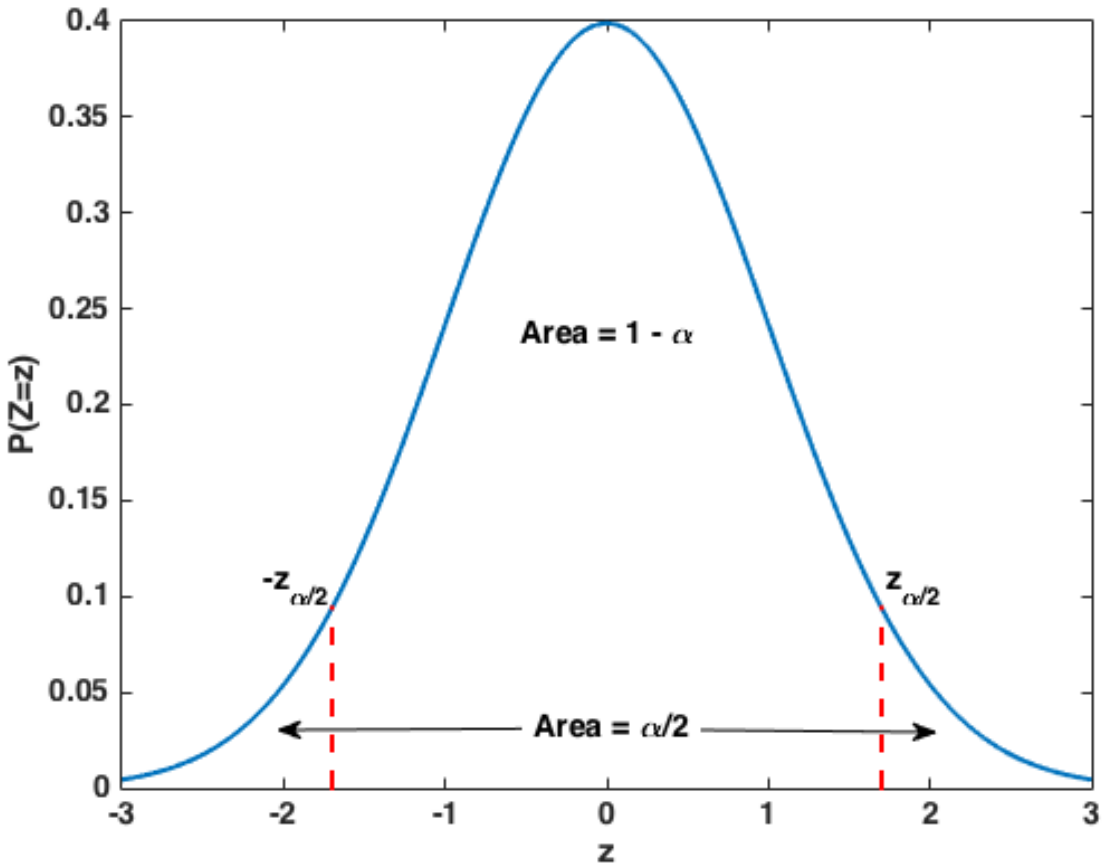
- For a 90% confidence interval, the significance level is 0.1

Constructing a Confidence Interval

When the population standard deviation σ is known, the confidence interval is determined as follows:

1. Compute the standard error of the mean $\sigma_{\bar{x}} = \sigma/\sqrt{n}$
2. Find the z-value $z_{\alpha/2}$ s.t. $P(Z \leq z_{\alpha/2}) = 1 - \alpha/2$
   - $Z \sim \mathcal{N}(0, 1)$
   - α is the level of significance
   - $1 - \alpha$ is the confidence level
   - $z_{\alpha/2}$ can be found by looking up the probability $1 - \alpha/2$ in a normal distribution table, `qnorm` in R, `scipy.stats.norm.ppf` in Python
3. Compute the margin of error $E = z_{\alpha/2}\sigma_{\bar{x}}$
4. Confidence interval $CI = [\bar{x} - E, \bar{x} + E]$

Meaning of $z_{\alpha/2}$



Example

Suppose that a signal with value μ is received with a value that is normally distributed around μ with variance 4

To reduce error, the signal is transmitted 10 times.

If the values received are $8.5, 9.5, 9.5, 7.5, 9, 8.5, 10.5, 11, 11$ and 7.5, then what is a 95% confidence interval for μ?

Example cont'd

1. Compute the sample mean $\bar{x} = 9.25$
2. Standard error of the mean

$$\sigma_{\bar{x}} = \sigma/\sqrt{n} = 2/\sqrt{10} = 0.6325$$

3. Using $\alpha = 0.05$, $1 - \alpha/2 = 0.975$, we get $z_{\alpha/2} = 1.96$
4. Margin of error

$$E = z_{\alpha/2}\sigma_{\bar{x}} = (1.96)(0.6325) = 1.24$$

5. Confidence interval

$$CI = [\bar{x} - E, \bar{x} + E] = [9.25 - 1.24, 9.25 + 1.24] = [8.01, 10.49]$$

Interpreting Confidence Intervals

Once the confidence interval is obtained, it is essential to interpret it correctly

Given a 90% confidence interval, it is not true that the population mean has a 90% probability of falling within the interval

There is a 90% probability that any given confidence interval from a random sample will contain the population mean

All confidence intervals for a given confidence level and sample size have the same width E, but the center is the sample mean, which can vary

Changing the Confidence Level

Significance level α represents the probability of erroneously concluding that the population mean is outside the confidence interval, when in fact it lies within the interval

As the confidence level $1 - \alpha$ increases, the significance level α decreases (since these two quantities must sum to one), which causes the z-score $z_{\alpha/2}$ to increase, and therefore the interval widens

Chance of erroneously concluding that the population mean is outside the confidence interval decreases

Law of Large Numbers

As the sample size n increases, the standard error of the mean decreases

Margin of error E decreases and the confidence interval shrinks

With a larger sample size, the sample mean more accurately approximates the population mean

Law of Large Numbers: As $n \to \infty$, the sample mean $\bar{x}$ converges to the population mean μ

Choosing the Sample Size for the Mean

Given a desired margin of error E, one can solve for the sample size n that would produce this value of E for the width of the interval

Rearranging the formulas for the confidence interval, we obtain

$$n = \left(\frac{\sigma}{\sigma_{\bar{x}}}\right)^2 = \left(\frac{\sigma z_{\alpha/2}}{E}\right)^2$$

Inversely proportional: as the margin of error E decreases, the sample size n increases

Large Samples: When σ is Unknown

If the population standard deviation σ is unknown, a confidence interval can be obtained by substituting the sample standard deviation s

Bootstrap Principle

Standard error of the mean:

$$\hat{\sigma}_{\bar{x}} = s/\sqrt{n}$$

Small Samples: When σ is Known

When the sample size n is considered small ($n < 30$), we cannot use the Central Limit Theorem to conclude that the sampling distribution of the mean is normal

We assume that the population itself is normal

When the population standard deviation σ is known, then we can proceed in the same way as for large samples

Small Samples: When σ is Unknown

When σ is unknown, we can substitute s for σ as is done for large samples, but to determine the margin of error E, instead of using the z-value $z_{\alpha/2}$ from the normal distribution, we use Student's t-distribution $t_{\alpha/2,n-1}$

This distribution, like the normal distribution, is bell-shaped and symmetric around the mean, and the area under the probability density curve is 1, but the shape of this curve depends on the degrees of freedom which is $n-1$

This is because there are n observations in the sample, but one degree of freedom is removed due to the mean

Student's t-distribution curve is flatter than the normal distribution curve, but it converges to a normal distribution as n increases

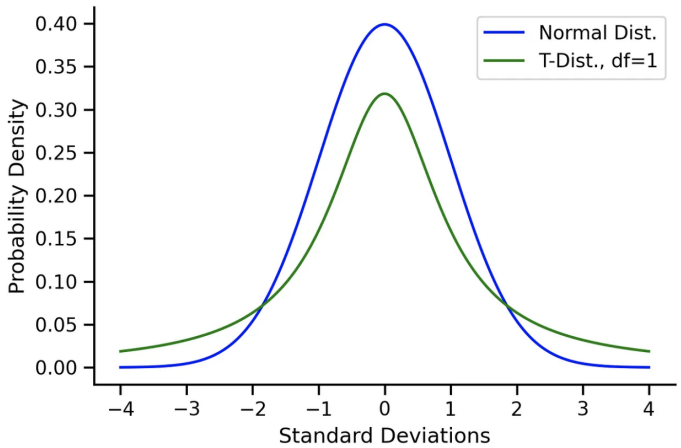
Using Student's t-distribution

Confidence interval:

$$CI = \left[\bar{x} - t_{\alpha/2,n-1}\hat{\sigma}_{\bar{x}},\ \bar{x} + t_{\alpha/2,n-1}\hat{\sigma}_{\bar{x}}\right], \quad \hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Value of $t_{\alpha/2,n-1}$ can be obtained by:

- Looking up the probability $1 - \alpha/2$ in a Student's t-distribution table
- `qt` in R
- `scipy.stats.t.ppf` in Python



Ref: `https://tjkyner.medium.com/the-normal-distribution-vs-students-t-distribution-322aa12ffd15`

Hypothesis Testing

Applications of inferential statistics in which a sample is used to determine, within a certain level of confidence, whether to reject a hypothesis about the population from which the sample was drawn

Prime example of how statistics is useful for acquiring insights into populations from raw data

A hypothesis is defined to be an assumption about a population parameter

Formulate hypotheses about whether a certain parameter is less than, equal to, or greater than a certain value, and then use confidence intervals to test whether these hypotheses should be rejected

Null and Alternative Hypotheses

For hypothesis testing, we use two hypotheses:

1. Null hypothesis: H_0 represents the "status quo". It states a belief about how a population parameter, e.g., the mean, compares to a specific value
2. Alternative hypothesis: H_1 is the opposite of H_0

Stating the Null and Alternative Hypotheses

For hypothesis testing to be as useful as possible, it is important to choose the alternative hypothesis H_1 wisely

The alternative hypothesis plays the role of the "research hypothesis", i.e., it corresponds to the position that the researcher wants to establish

Example

Suppose that a brand of lightbulbs has a mean lifetime of 2000 hours, but an improvement has been made to their design that may extend their lifetime

1. Null hypothesis: $H_0 : \mu \leq 2000$
2. Alternative hypothesis: $H_1 : \mu > 2000$

If it is determined that H_0 should be rejected, there is evidence to support the claim that the newly designed lightbulbs do have a longer lifetime

Process of Hypothesis Testing

1. Determine a rejection region of the sampling distribution for the parameter, e.g., mean featured in H_0
2. Check whether an appropriate test statistic e.g., sample mean falls within the rejection region
3. If so, we choose to reject H_0, and conclude that there is sufficient evidence to support the claim made by H_1

   Otherwise, we choose not to reject H_0, and conclude that there is not sufficient evidence to support the claim made by H_1

Hypothesis test does not provide enough evidence to accept H_0; we are only concerned with whether to reject it

Type I and Type II Errors

Because of the reliance on a sample, it is possible for the conclusion of a hypothesis test to be erroneous

1. Type I error: reject H_0 when it is true
   - Often due to a sampling error
   - Probability is the level of significance α used to construct the confidence interval for the hypothesis test
2. Type II error: not reject H_0 when it is false
   - Probability of error is denoted by β
   - For a fixed sample size, β decreases as α increases

Probability of both errors can be decreased by increasing the sample size

Two-Tail Hypothesis Testing

Two-tail hypothesis test: Null hypothesis H_0 is a statement of equality

Example: null hypothesis for the mean $H_0 : \mu = \mu_0$, for some value of μ_0

Role of Confidence Intervals

First choose the significance level α, based on what is considered an acceptable probability of making a Type I error

Construct a confidence interval around μ_0:

$$[\mu_0 - z_{\alpha/2}\sigma_{\bar{x}},\ \mu_0 + z_{\alpha/2}\sigma_{\bar{x}}]$$

If the sample mean $\bar{x}$ falls within this confidence interval, then we do not reject H_0

Otherwise, we say that $\bar{x}$ falls within the rejection region (subset of the real numbers outside of the confidence interval) and we reject H_0

Test Statistic

By rearranging algebraically, we obtain the equivalent condition that we do not reject H_0 if the test statistic

$$z^* = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

satisfies

$$-z_{\alpha/2} \leq z^* \leq z_{\alpha/2}$$

since

$$\mu_0 - z_{\alpha/2}\sigma_{\bar{x}} \leq \bar{x} \leq \mu_0 + z_{\alpha/2}\sigma_{\bar{x}}$$

One-Tail Hypothesis Testing

One-tail hypothesis test: Null hypothesis H_0 is an inequality

Example: a null hypothesis for the mean $H_0 : \mu \leq \mu_0$ or $H_0 : \mu \geq \mu_0$

One-sided Confidence Intervals

As with the two-tail test, we first choose the significance level α. Then construct a one-sided confidence interval

For the null hypothesis $H_0 : \mu \leq \mu_0$, the interval is

$$(-\infty, \mu_0 + z_\alpha \sigma_{\bar{x}}]$$

If the sample mean $\bar{x}$ falls within this confidence interval (that is, $\bar{x} \leq \mu_0 + z_\alpha \sigma_{\bar{x}}$), then we do not reject H_0

Otherwise, if $\bar{x} > \mu_0 + z_\alpha \sigma_{\bar{x}}$, then $\bar{x}$ falls within the rejection region and we reject H_0

Test Statistics with One-Tail Tests

Equivalently, we do not reject H_0 if the test statistic satisfies

$$z^* \leq z_\alpha$$

Similarly, if the null hypothesis is $H_0 : \mu \geq \mu_0$, we do not reject H_0 if

$$z^* \geq -z_\alpha$$

Note that one-tail hypothesis testing uses the same test statistic as in the two-tail case, but it is compared to different values

Hypothesis Testing with One Sample

General idea is the same: a confidence interval needs to be constructed around the value that is compared to the parameter in H_0

Outside this interval lies the rejection region; if the test statistic falls within the rejection region, then H_0 is rejected

The differences between scenarios relate to the various parameters used to construct the confidence interval

Testing for the Mean: Large Sample

First, we consider hypothesis testing for the case in which the parameter of interest is the mean, and the sample is large ($n \geq 30$)

By the Central Limit Theorem, the sampling distribution of the mean is well approximated by a normal distribution

We consider both one-tail hypothesis tests ($H_0 : \mu \geq \mu_0$ or $H_0 : \mu \leq \mu_0$) and two-tail tests ($H_0 : \mu = \mu_0$)

When σ is Known

When the population standard deviation σ is known, then the appropriate test statistic is:

$$z^* = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}}$$

where $\sigma_{\bar{x}} = \sigma/\sqrt{n}$ is the standard error of the mean

Example

A commercial hatchery grows salmon whose weights are normally distributed with a standard deviation of 1.2 pounds ($\sigma = 1.2$)

The hatchery claims that the mean weight is at least 7.6 pounds ($H_0 : \mu \geq 7.6$)

Suppose a random sample of 40 fish yields an average weight of 7.2 pounds ($n = 40, \bar{x} = 7.2$)

Is this strong enough evidence to reject the hatchery's claim at the 5% level of significance? ($\alpha = 0.05$)

Testing the Hypothesis

Null hypothesis $H_0 : \mu \geq 7.6$ and alternative hypothesis $H_1 : \mu < 7.6$

Standard error:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} = \frac{1.2}{\sqrt{40}} = 0.1897$$

Test statistic:

$$z^* = \frac{\bar{x} - \mu_0}{\sigma_{\bar{x}}} = \frac{7.2 - 7.6}{0.1897} = -2.1082$$

We then compare this value to $-z_\alpha = -z_{0.05} = -1.6449$

Because $z^* < -z_\alpha$, the test statistic falls within the rejection region and therefore we reject H_0 and conclude that the hatchery's claim does not have merit

When σ is Unknown

Substitute s, the sample standard deviation, for σ and proceed as before

Because the sample is large, it is assumed that s is a reasonably accurate approximation for σ

Test statistic:

$$z^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

Example

Twenty years ago, male students at a high school could do an average of 24 pushups in 60 seconds ($\mu = 24$)

To determine whether this is still true today, a sample of 50 male students is chosen ($n = 50$)

The sample mean is 22.5 pushups and the sample standard deviation is 3.1 ($\bar{x} = 22.5, s = 3.1$)

Can we conclude that the mean is no longer 24 at the 5% level of significance? ($H_0 : \mu = 24$, $\alpha = 0.05$)

Example cont'd

Null hypothesis $H_0 : \mu = 24$ and the alternative hypothesis $H_1 : \mu \neq 24$

Since the population standard deviation is unknown, but the sample is sufficiently large, we use the sample standard deviation instead

Standard error:

$$\hat{\sigma}_{\bar{x}} = \frac{s}{\sqrt{n}} = \frac{3.1}{\sqrt{50}} = 0.4384$$

Example cont'd

Test statistic:

$$z^* = \frac{\bar{x} - \mu_0}{\hat{\sigma}_{\bar{x}}} = \frac{22.5 - 24}{0.4384} = -3.4215$$

Because this is a two-tail test, we compare z^* to $z_{\alpha/2} = z_{0.025} = 1.96$

We have $|z^*| > z_{\alpha/2}$, which means z^* falls within the rejection region

Therefore, we reject H_0 and conclude that the mean is no longer 24

Role of α

It can be seen from examination of a normal distribution table that as α increases, z_α (or $z_{\alpha/2}$) decreases

- $P(Z > z_\alpha) = \alpha$
- $P(Z \leq z_\alpha) = 1 - \alpha$

Test statistic is less likely to fall within the appropriate confidence interval for the hypothesis test; that is, it is more likely that H_0 will be rejected

Role of α

Considering that the alternative hypothesis H_1 is generally the one that supports a position that a researcher is trying to establish, it is in the researcher's interest that H_0 be rejected

As such, they can help their cause by choosing a larger value of α, which corresponds to a lower confidence level $1 - \alpha$

This is an important ethical consideration for a statistician and underscores the importance of knowing the parameters used in any statistical analysis that is used to support a particular position

The smaller the value of α, the more confidence (pun intended) one can have in the result of a hypothesis test

How to choose α: p-values

Important to avoid a Type I error (rejecting H_0 when it is actually valid)

Probability of making this error is the level of significance α

Helpful to have some guidance in choosing α

p-value: smallest value of significance at which H_0 will be rejected assuming it is true

One-Tail Test $H_0 : \mu \leq \mu_0$

Null hypothesis $H_0 : \mu \leq \mu_0$ and alternative hypothesis $H_1 : \mu > \mu_0$

H_0 is rejected if

$$z^* > z_\alpha$$

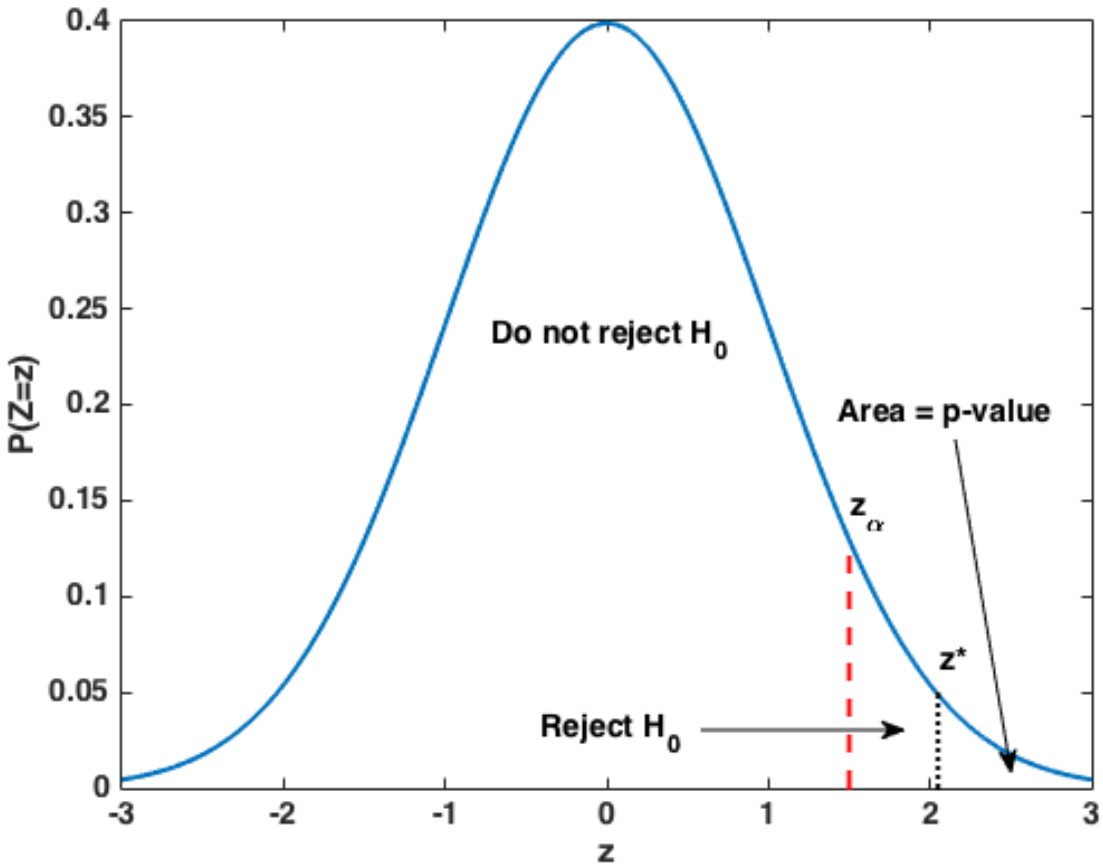
Because z_α satisfies $P(Z > z_\alpha) = \alpha$, H_0 is rejected if

$$P\left(Z > z^*\right) < \alpha$$

p-value for this hypothesis test:

$$P\left(Z > z^*\right)$$

What is the *p*-value?



One-Tail Test $H_0 : \mu \geq \mu_0$

Null hypothesis $H_0 : \mu \geq \mu_0$ and alternative hypothesis $H_1 : \mu < \mu_0$

H_0 is rejected if

$$z^* < -z_\alpha$$

Because $P(Z \leq -z_\alpha) = \alpha$, H_0 is rejected if

$$P\left(Z \leq z^*\right) < \alpha$$

p-value for this hypothesis test:

$$P(Z \leq z^*)$$

Two-Tail Tests

Finding a p-value for a two-tail test is similar to the one-tail case

Null hypothesis $H_0 : \mu = \mu_0$ is rejected if

$$|z^*| > z_{\alpha/2}$$

Because $z_{\alpha/2}$ satisfies $P(|Z| > z_{\alpha/2}) = \alpha$, H_0 is rejected if

$$P\left(|Z| > |z^*|\right) < \alpha$$

Due to the symmetry of the normal distribution, this condition is equivalent to

$$P\left(Z > |z^*|\right) < \alpha/2$$

p-value for a two-tail test is twice the p-value of the one-tail test

Interpreting p-values

Null hypothesis is rejected if the p-value is smaller than the significance level α

Example: p-value 0.02 means that H_0 is rejected at the 95% confidence level, but not at the 99% level

Researchers like small p-values as they imply statistically significant results

Manipulating p-values

Unfortunately, this means researchers may (intentionally or otherwise) improperly influence p-values to drive them toward zero

For example, this can be achieved by measuring many quantities on a small sample, which almost guarantees a statistically significant result

For more information:
"I Fooled Millions Into Thinking Chocolate Helps Weight Loss. Here's How." by John Bohannon, posted on `io9.com`

Remember the advice of Ronald Fisher, who introduced p-values: a small p-value does not prove a research hypothesis true! It only means the evidence is worth a second look

Testing for the Mean: Small Sample

When the sample size is small ($n < 30$), we can no longer rely on the Central Limit Theorem and automatically treat the sampling distribution of the mean as a normal distribution

We must assume that the population itself is normally distributed for our hypothesis testing procedures to remain valid

Small Samples: When σ is Known

When the population standard deviation σ is known, then we can proceed with hypothesis testing as before provided that the population is in fact normally distributed

If this is not the case, then the result of a hypothesis test may be unreliable.

Small Samples: When σ is Unknown

Assume population is normally distributed

Need alternative approach to computing the threshold against which to compare the test statistic

Need an alternative value that plays the role of z_α in a one-tail test or $z_{\alpha/2}$ in a two-tail test

Use Student's t-distribution, as we did when constructing confidence intervals using small samples with σ unknown

Using Student's t-distribution

As in the case of a large sample with σ unknown, our test statistic is

$$t^* = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$$

This test statistic is compared to a value of Student's t-distribution with $n-1$ degrees of freedom, where n is the sample size

For a given significance level α, $t_{\alpha,n-1}$ is the t-value such that $P(T_{n-1} > t_{\alpha,n-1}) = \alpha$, where T_{n-1} is a random variable that follows Student's t-distribution with $n-1$ degrees of freedom

When to Reject?

For a one-tail test, with null hypothesis

- $H_0 : \mu \leq \mu_0$, reject H_0 if $t^* > t_{\alpha,n-1}$ and do not reject H_0 otherwise
- $H_0 : \mu \geq \mu_0$, reject H_0 if $t^* < -t_{\alpha,n-1}$ and do not reject H_0 otherwise

For a two-tail test with null hypothesis

- $H_0 : \mu = \mu_0$, reject H_0 if $|t^*| > t_{\alpha/2,n-1}$, and do not reject H_0 if $|t^*| \leq t_{\alpha/2,n-1}$

Hypothesis Testing with Two Samples

Testing hypotheses with characteristics of two populations

Examples:

- investigating differences in test scores between males and females
- comparison of long-life vs standard light bulbs
- average selling prices of homes in different areas

Sampling Distribution for the Difference of Means

Two-sample hypothesis testing can be used to compare means

Sampling distribution for the difference in means: probability of observing various intervals for the difference between two sample means

Standard error of the difference:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

where the first sample of size n_1 has standard deviation σ_1 and the second sample of size n_2 has standard deviation σ_2

Testing for Difference of Means: Large Samples

Assume that the two samples are independent

If the sample sizes n_1 and n_2 are large, then we assume the sampling distribution of the difference of means is normal

Test statistic:

$$z^* = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}}$$

Example

Two new methods of producing a tire:

1. $n_1 = 40$ tires are tested at location A that have a mean lifetime of $\bar{x}_1 = 40,000$ miles with standard deviation of $\sigma_1 = 4,000$ miles
2. $n_2 = 50$ tires are tested at location B that have a mean lifetime of $\bar{x}_2 = 42,000$ miles with standard deviation of $\sigma_2 = 5,000$ miles

Test the hypothesis that both methods produce tires with the same average lifetimes at the 5% significance level ($\alpha = 0.05$)

Example cont'd

Null hypothesis $H_0 : \mu_1 = \mu_2$ and the alternative hypothesis $H_1 : \mu_1 \neq \mu_2$

Standard error:

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}} = \sqrt{\frac{4000^2}{40} + \frac{5000^2}{50}} = 948.6833$$

Test statistic:

$$z^* = \frac{\bar{x}_1 - \bar{x}_2}{\sigma_{\bar{x}_1 - \bar{x}_2}} = \frac{40,000 - 42,000}{948.6833} = -2.1082$$

Compare this against $z_{\alpha/2} = z_{0.05/2} = 1.96$

Example cont'd

Since $|z^*| > z_{\alpha/2}$, we reject the null hypothesis and conclude that the two methods produce tires with statistically different average lifetimes

p-value:

$$P(|Z| > |-2.1082|) = 2P(Z > 2.1082) = 2(1 - P(Z \leq 2.1082)) = 0.035$$

Null hypothesis is rejected at any significance level above 3.5%

Testing for Difference of Means: Unknown Variance

If the standard deviations are unknown, then we must use Student's t-distribution

When the sample sizes are small ($n_1, n_2 < 30$) we assume that the populations are normally distributed

For now, we assume that the samples are independent

This kind of hypothesis test is called an unpaired t-test

Equal Standard Deviations

When the population standard deviations are unknown but assumed to be equal, we use the sample standard deviations to obtain a pooled estimate of standard deviation:

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

Number of degrees of freedom:

$$d.f. = n_1 + n_2 - 2$$

Degrees of freedom of the samples, $n_1 - 1$ and $n_2 - 1$, are added to obtain the degrees of freedom to be used for the test

Equal Standard Deviations cont'd

Standard error of the difference of means:

$$\hat{\sigma}_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$$

Test statistic:

$$t^* = \frac{\bar{d} - d_0}{\hat{\sigma}_{\bar{d}}}$$

where d represents $x_1 - x_2$ and d_0 is the value against which we are testing

Unequal Standard Deviations

When the standard deviations are unequal, we perform an unpooled test

Standard error of the difference of means:

$$\hat{\sigma}_{\bar{x}_1-\bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Number of degrees of freedom:

$$d.f. = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\frac{\left(\frac{s_1^2}{n_1}\right)^2}{n_1 - 1} + \frac{\left(\frac{s_2^2}{n_2}\right)^2}{n_2 - 1}}$$

which must be rounded to an integer

Testing for Difference of Means: Dependent Samples

Suppose that the two samples are dependent

Example: testing the average weight loss of a group of individuals, in which each person's original weight is paired with their current weight

Use Student's t-distribution for a paired t-test

Example

Suppose that a group of 10 patients is given medication that is intended to lower their cholesterol

Their cholesterol is tested before and after being given the medication, and they are found to have their cholesterol level lowered by an average of $\bar{d} = 10$ mg/dL, with a sample standard deviation of $s_d = 8$ mg/dL

If we test at the 1% significance level ($\alpha = 0.01$), is the reduction in cholesterol level statistically significant?

Example cont'd

Null hypothesis: Medication does not help $H_0 : \mu \leq 0$
Alternative hypothesis: $H_1 : \mu > 0$
μ: mean reduction in cholesterol level

Standard Error:

$$\hat{\sigma}_{\bar{d}} = \frac{s_d}{\sqrt{n}} = \frac{8}{\sqrt{10}} = 2.5298$$

Test statistic:

$$t^* = \frac{\bar{d} - 0}{\hat{\sigma}_{\bar{d}}} = \frac{10}{2.5298} = 3.9528$$

Compare this value to $t_{\alpha,n-1} = t_{0.01,9} = 2.8214$

Because $t^* > t_{\alpha,n-1}$, we reject H_0 and conclude that the reduction in cholesterol level is statistically significant

Summary

There are several hypothesis tests for different situations, and various ways to conduct the test that are mathematically equivalent

To help keep track of these situations:

- Test statistic is always the difference between the value of the variable being tested (e.g. the sample mean) and the value it's being tested against (e.g. μ_0 if the null hypothesis is $H_0 : \mu = \mu_0$) divided by the standard error
- Use Student's t-distribution if the variances are unknown and sample standard deviations to obtain the standard error

Standard Errors

Characteristic	Scenario	Standard Error
μ	Variance known	$\sigma_{\bar{x}} = \sigma/\sqrt{n}$
μ	Variance unknown	$\hat{\sigma}_{\bar{x}} = s/\sqrt{n}$
p	n sample	$\sigma_p = \sqrt{\frac{p_0(1-p_0)}{n}}$
$\mu_1 - \mu_2$	n_i large, σ_i known, independent samples	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$

Standard Errors cont'd

Characteristic	Scenario	Standard Error
$\mu_1 - \mu_2$	n_i large, σ_i unknown, $\sigma_1 = \sigma_2$, independent	$\sigma_{\bar{x}_1 - \bar{x}_2} = s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}$ $s_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$
$\mu_1 - \mu_2$	n_i large, σ_i unknown, $\sigma_1 \neq \sigma_2$, independent	$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$
$d = \mu_1 - \mu_2$	n large, σ unknown, dependent samples	$\hat{\sigma}_{\bar{d}} = s_d / \sqrt{n}$

Introduction

In this section, we will learn how to use correlation and regression to gain some insight into the nature of the relationship between two variables

Independent and Dependent Variables

Independent variable: x
Dependent variable: y

x serves as the "input" and y serves as the "output"

Mathematically, y is a function of x, meaning that y is determined from x in some systematic way

For each value of x, there is only one value of y, whereas one value of y can correspond to more than one value of x

Correlation

Correlation measures the strength and direction of the relationship between x and y

Types of correlation are:

- positive linear correlation: as x increases, y increases linearly
- negative linear correlation: as x increases, y decreases linearly
- nonlinear correlation: clear relationship between x and y, but the dependence of y on x cannot be described graphically using a straight line
- no correlation: no clear relationship between x and y

We will focus on linear correlation

Correlation Coefficient

To determine the correlation between two variables x and y, for which we have n observations (x_i, y_i), we compute the correlation coefficient:

$$r = \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{\sqrt{\left[n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2\right]\left[n\sum_{i=1}^{n} y_i^2 - \left(\sum_{i=1}^{n} y_i\right)^2\right]}}$$

Geometrically, r is the cosine of the angle between the vector of x-values and the vector of y-values with their means subtracted

It follows from this interpretation that $|r| \leq 1$

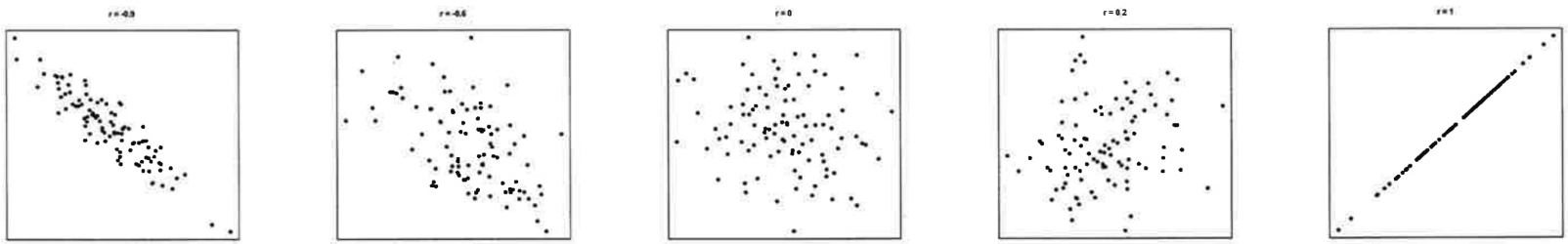
Interpretation

Sign gives the direction

- If $r > 0$, then x and y have a positive linear correlation
- If $r < 0$, then x and y have a negative linear correlation
- If $r = 0$, then there is no correlation between x and y

Magnitude gives the strength

- In extreme cases, $r = \pm 1$, we have $y = cx$ for some constant c that is positive ($r = 1$) or negative ($r = -1$)
- Accuracy of this prediction depends on $|r|$; if r is nearly zero, prediction is not likely to be reliable



Testing the Significance of r

Suppose we have determined that x and y are linearly correlated, based on the value of the correlation coefficient r obtained from a sample

How do we know whether a similar correlation applies to the entire population?

Perform a hypothesis test on the population correlation coefficient p

Test whether p is nonzero, use a two-tail test with null hypothesis $H_0 : p = 0$ and alternative hypothesis $H_1 : p \neq 0$

Test for a positive linear correlation, use a one-tail test with null hypothesis $H_0 : p \leq 0$ and alternative hypothesis $H_1 : p > 0$

Performing the Test

For this test, we use Student's t-distribution with $d.f. = n - 2$

Standard error of correlation coefficient:

$$s_r = \sqrt{\frac{1 - r^2}{n - 2}}$$

Test statistic:

$$t^* = \frac{r}{s_r}$$

For the two-tail test with $H_0 : p = 0$, we reject H_0 and conclude that x and y are linearly correlated if $|t^*| > t_{\alpha/2,n-2}$

For the one-tail test with $H_0 : p \leq 0$, we reject H_0 and conclude that x and y have a positive linear correlation if $t^* > t_{\alpha,n-2}$

Correlation vs. Causation

Always keep in mind: correlation does not imply causation!

Meaning: it often occurs that variables exhibit a correlation with one another even though there is no influence whatsoever

Even if there is a causal relationship, it's not always clear which is the cause and which is the effect!

Reverse Causality

"Effect" of Course Signals on student retention at Purdue University

Purdue developed Course Signals to use analytics to alert faculty and staff to potential problems for students

Purdue claimed that when students took at least two courses that used Course Signals, retention improved by 21%!

This conclusion was supported by appropriate data, so what could be the problem?

Look for Anomalies!

It was observed from the data that taking two Course Signals courses greatly improved retention whereas taking only one did not help at all

Also, an initial bump in retention rate quickly faded after Course Signals had been in use for a few years

What the data was really showing was that students were taking more Course Signals courses because they were taking more courses overall (that is, they did not control for freshmen dropping out early)

In other words, it was retention that led to increased use of Course Signals, not the other way around!

Reference: "What the Course Signals 'Kerfuffle' is About, and What it Means to You" by Michael Caulfield, posted at `educause.edu`

Causal Inference

Given that two variables are correlated, the ideal approach to establishing causation is to understand the mechanism by which it acts

Failing that, another approach, if less effective, is to perform a controlled intervention study

Establishing causation based solely on observations is much less reliable, but more broadly applicable

In fact, this is impossible without making assumptions about the data

Reference: Max Planck Institute

Linear Regression

If x and y are found to be linearly correlated, then we use linear regression to find the straight line that best fits the ordered pairs $\{(x_i, y_i)\}_{i=1}^{n}$

Equation of this line:

$$\hat{y} = a + bx$$

where $\hat{y}$ is the predicted value of y obtained from x

The y-intercept a and slope b need to be determined

Least Squares Method

To find the values of a and b such that the line $\hat{y} = a + bx$ best fits the sample data, we use the least squares method

Compute a and b to minimize:

$$\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = \sum_{i=1}^{n}(y_i - a - bx_i)^2$$

The name of the method comes from the fact that we are trying to minimize a sum of squares of the deviations between y and $\hat{y}$

The line $\hat{y} = a + bx$ that minimizes this sum of squares and best fits the data is the regression line

Solving the Least Squares Problem

Minimizing coefficients:

$$
\begin{aligned}
a &= \bar{y} - b\bar{x} \\
b &= \frac{n\sum_{i=1}^{n} x_i y_i - \sum_{i=1}^{n} x_i \sum_{i=1}^{n} y_i}{n\sum_{i=1}^{n} x_i^2 - \left(\sum_{i=1}^{n} x_i\right)^2}
\end{aligned}
$$

where $\bar{x}$ and $\bar{y}$ are the sample means $\bar{x} = \sum_{i=1}^{n} x_i, \quad \bar{y} = \sum_{i=1}^{n} y_i$

Connection to correlation coefficient r

b is closely related to the correlation coefficient r

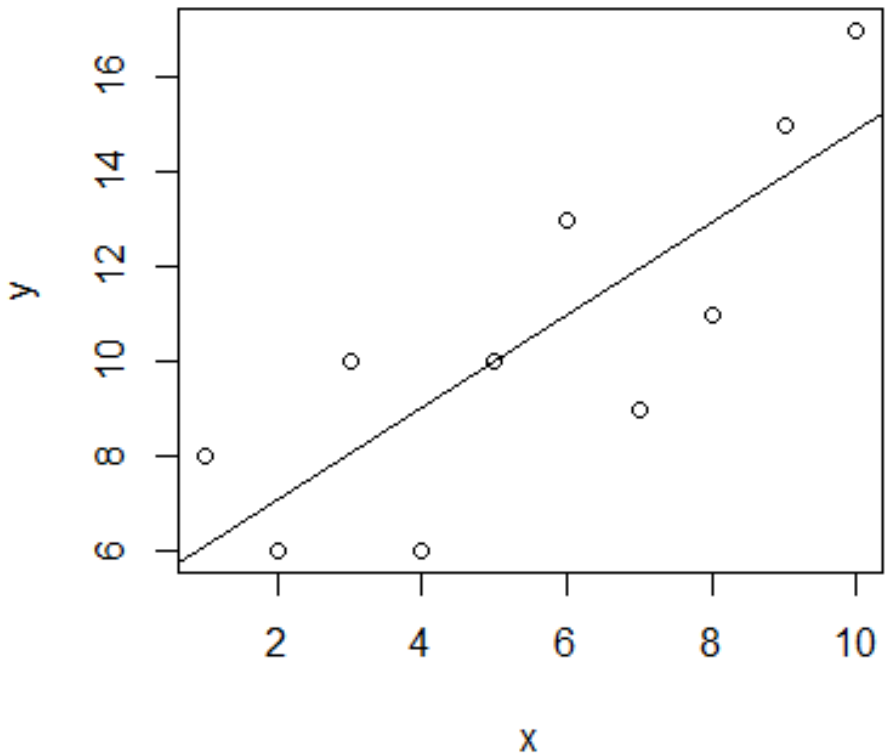
The formulas have the same numerator:

$$b = r\frac{s_y}{s_x}$$

Slope is positive if and only if the correlation coefficient indicates that x and y have a positive linear correlation

Plot of Regression Line

It is a coincidence that the regression line happens to pass through one of the points; in general this does not happen, as the goal of the least squares method is to minimize the distance between all of the predicted y-values and observed y-values



Confidence Interval for the Regression Line

To measure how well the regression line fits the data, we can construct a confidence interval

Standard error of the estimate:

$$s_e = \sqrt{\frac{\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}{n-2}}$$

Measures amount of dispersion of the observations around regression line

Smaller s_e is the closer the points are to the regression line

Similarity between this formula and the sample standard deviation; the number of degrees of freedom is $n-2$ since two degrees of freedom are taken away by the coefficients a and b of the regression line

Testing the Slope of the Regression Line

Need to determine whether the slope b of the regression line is indicative of the slope β for the population

Perform a hypothesis test

Null hypothesis $H_0 : \beta = \beta_0$ and $H_1 : \beta \neq \beta_0$ for a two-tail test

If $\beta_0 = 0$, then we are testing whether there is any linear relationship between x and y, and rejection of H_0 would imply that this is the case

Standard Error of the Slope

Standard error of slope:

$$s_b = \frac{s_e}{\sqrt{\sum_{i=1}^{n} nx_i^2 - \bar{x}^2}}$$

where s_e is the standard error of the estimate

s_b is the standard deviation in the y-values divided by $\sqrt{n}$ times the standard deviation of the x-values

Intuitively makes sense because we are testing the slope which is the change in y divided by the change in x

Test Statistic

As with the test of the correlation coefficient, we use Student's t-distribution to determine the critical value

Test statistic:

$$t^* = \frac{b - \beta_0}{s_b}$$

Compared to the critical value $t_{\alpha/2,n-2}$, the t-value satisfying $P(|T_{n-2}| > t_{\alpha/2,n-2}) = \alpha/2$

If $|t^*| > t_{\alpha/2,n-2}$, then we reject H_0 and conclude $\beta \neq \beta_0$

If $\beta_0 = 0$, then our conclusion is that x and y are linearly correlated

Assumptions

For the least squares method to be valid, we need to make the following assumptions:

- Individual differences between y_i and $\hat{y}_i$, $i = 1, 2, \ldots, n$, are independent of one another
- Observed values of y are normally distributed around $\hat{y}$
- Variation of y around the regression line is equal for all values of x

Other Issues to Consider

- Avoid predicting y by **extrapolation**, i.e., at x-values outside the range of x-values that were used for the regression: the linear relationship often breaks down outside a certain range
- Beware of data that are summaries (e.g., averages of some data). Those are less variable than individual observations and correlations between averages tend to overstate the strength of the relationship
- Regression analyses often report 'R-squared': $R^2 = r^2$. It gives the fraction of the variation in the y-values that is explained by the regression line. (So $1 - r^2$ is the fraction of the variation in the y-values that is left in the residuals)

Maximum Likelihood Estimation (MLE)

Let $x_1, x_2, \ldots, x_n$ be a sample of n i.i.d (independent and identically distributed) observations, coming from an unknown distribution with probability distribution function of the form $f(x, \theta)$

The method of maximum likelihood is used to obtain an estimate $\hat{\theta}$ of the unknown parameter θ

Observations are independent:

$$f(x_1 \cap x_2 \cap \cdots \cap x_n|\theta) = f(x_1|\theta)f(x_2|\theta)\cdots f(x_n|\theta)$$

The maximum likelihood estimator (MLE) is the value of $\hat{\theta}$ that maximizes the average log-likelihood

$$\hat{\ell} = \frac{1}{n}\sum_{i=1}^{n} \log f(x_i|\theta)$$

Example

Let the n observations be coin flips of an unfair coin, and let h be the number of heads. These flips follow a binomial distribution

$$f(X = h|\theta) = \left(\begin{array}{c} n \\ h \end{array} \right) \theta^h (1-\theta)^{n-h}$$

with unknown probability of success θ

The MLE $\hat{\theta}$ maximizes

$$\frac{1}{n} \log \left(\begin{array}{c} n \\ h \end{array} \right) \theta^h (1-\theta)^{n-h} = \frac{1}{n} \left[\log \left(\begin{array}{c} n \\ h \end{array} \right) + h \log \theta + (n-h) \log(1-\theta) \right]$$

Take derivative and set to 0: maximized at $\hat{\theta} = h/n$

Main Takeaways

- Statistics is a part of everyday life
- Statistics provides mathematical fundamentals and foundations for modern day machine learning
- Visualize and understand your data as a first step in solving your problem as data scientists

Thank you!

Additional Related Workshops

- Linear Algebra
- Intro to Python/Python for DataSc
- Intro to ML and related courses
- Data Visualization

Questions:

- Danielle: `dcmaddix@gmail.com`
- Eden: `edenl@stanford.edu`
- Piazza: `https://piazza.com/class/lj76htp21lk2o`