

Course Topics

1. Elementary Probability
2. Conditional Probability
3. Discrete Random Variables
4. Expectation, Variance and Moments
5. The Pascal, Negative Binomial and Poisson Distributions
6. Continuous Random Variables
7. The Normal Distribution
8. Multivariate Random Variables
9. The Hypergeometric Distribution
10. Transformation of Random Variables and Reliability
11. Samples and Data
12. Parameter Estimation
13. Interval Estimation
14. The Fisher Test

Course Topics

15. Neyman-Pearson Decision Theory
16. Null Hypothesis Significance Testing
17. Single Sample Tests for the Mean and Variance
18. Non-Parametric Single Sample Tests for the Median
19. Inferences on Proportions
20. Comparison of Two Variances
21. Comparison of Two Means
22. Non-Parametric Comparisons; Paired Tests and Correlation
23. Categorical Data
24. Simple Linear Regression I: Basic Model and Inferences
25. Simple Linear Regression II: Predictions and Model Analysis
26. Multiple Linear Regression I: Basic Model
27. Multiple Linear Regression II: Inferences on the Model
28. Multiple Linear Regression III: Finding the Right Model

Elementary Probability

Games of chance have a long history...



Cubical Die from Tepe Gawra. Photo of Object 31-52-309 of the Penn Museum. Online: <https://www.penn.museum/collections/object/332432>. Described in Brown, W. N. "Indian Games of Pachisi, Chaupar, and Chausar". Expedition: The Magazine of the University of Pennsylvania Museum of Archaeology and Anthropology. Philadelphia: The University Museum. 1964. Vol. 6, no. 3. Pages 32-35.

... but probability in mathematics does not. Why?

- ▶ **Platonism:** real-life objects are imperfect representations of ideal “Platonic Forms”. A six-sided die is a representation of an ideal cube.
- ▶ But randomness appears tied to real-life processes - no platonic form.
- ▶ Randomness was not considered amenable to mathematics.



Detail of the “School of Athens” by Rafael. Wall Fresco in the Vatican, Stanza della Segnatura. 1509. File:Raffael 058.jpg. (2019, September 1). Wikimedia Commons, the free media repository.

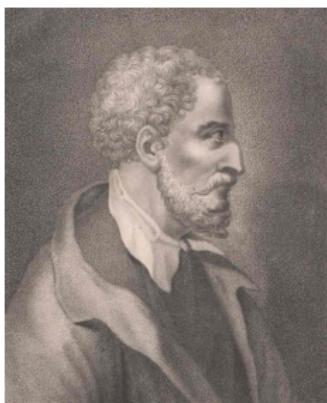
... but probability in mathematics does not. Why?

- ▶ ***Divination:*** randomness was often used to predict the future.
- ▶ Predicting randomness = interfering with the will of the gods.



Runestones. Online: <https://www.needpix.com/photo/download/1235148/divination-background-krupnyj-plan-the-consignment-a-few-runes-stones-scrying-stones-bone>.

Girolamo Cardano (1501-1576)



Girolamo Cardano (1501-1576). Etching by Carl Mayer. Dated 1813/1863 Online: <https://picryl.com/media/cardano-girolamo-1f9a09>.

- ▶ Invented cardan shaft
- ▶ Published solutions to cubic and quartic equations
- ▶ First systematic use of negative numbers in Europe; acknowledged imaginary numbers
- ▶ Heavy gambler, known to be short of money
- ▶ Published first systematic treatment of probability

1.1. **Cardano's Principle.** Let A be a random outcome of an experiment that may proceed in various ways. Assume each of these ways is equally likely. Then the probability $P[A]$ of the outcome A is

$$P[A] = \frac{\text{number of ways leading to outcome } A}{\text{number of ways the experiment can proceed.}}$$

Two Die Rolls

It is clear that the probability is a real number between 0 and 1.

1.2. Example. Two six-sided dice are rolled. Both are fair dice and have equal probability of returning any given number.

What is the probability that the sum of the results is 11 or 12?

There are six possible results for the first die and 6 possible results for the second die, so there are a total of $6 \cdot 6 = 36$ **possible outcomes**.

The outcomes that give a result of 11 or 12, writing the outcomes as (first die, second die), are:

- ▶ outcome (6, 6) gives the sum 12;
- ▶ outcome (5, 6) gives the sum 11;
- ▶ outcome (6, 5) gives the sum 11;
- ▶ all other outcomes will give a sum of 10 or less.

Two Die Rolls

The probability that the sum of the results is at least 11 is

$$\frac{\text{number of outcomes leading to a sum of 11 or 12}}{\text{number of possible outcomes}} = \frac{3}{36} = \frac{1}{12}.$$

When applying Cardano's principle, it is crucial that all outcomes are equally likely!

Tossing a Coin 10 Times

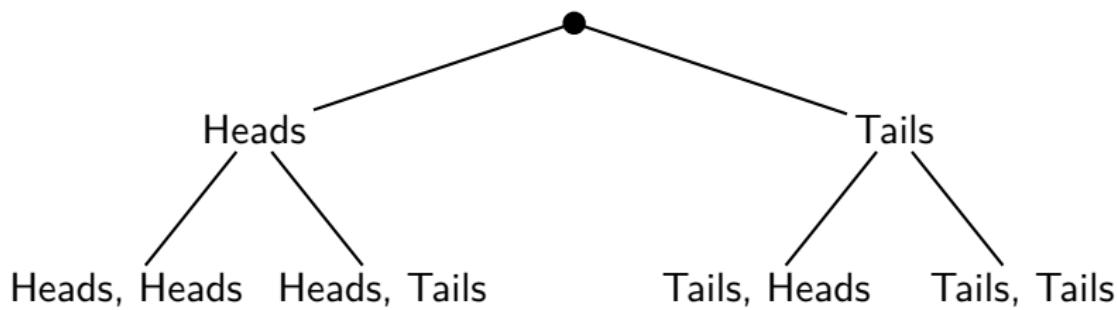
A fair coin is tossed ten times and the result of “heads” (h) or “tails” (t) is recorded each time. Which of the following sequences of results is the most likely:

- (a) $h, h, t, h, t, t, t, h, h, t;$
- (b) $h, h, h, h, h, t, t, t, t, t;$
- (c) $h, h, h, h, h, h, h, h, h, h;$
- (d) None of them (they are all equally likely).

Two Coin Tosses

1.3. Example. A fair coin is tossed twice and the result of “heads” (h) or “tails” (t) is recorded each time. What is the probability of obtaining at least one head?

We use a tree diagram to visualize the possible outcomes:



We see that

$$P[\text{at least 1 Head}] = \frac{3}{4}.$$

D'Alembert's Error



Portrait de Jean Le Rond d'Alembert, de la Tour, Maurice Quentin, 1753. Painting. Musée de Louvre, Paris. File:Alembert.jpg. (2020, January 29). Wikimedia Commons, the free media repository.

D'Alembert asserted in 1754 that it is erroneous to consider the four cases (h, h) , (h, t) , (t, h) , (t, t) since the experiment can be stopped immediately if heads comes up on the first toss.

Therefore, he claimed, there are only three outcomes:

- ▶ heads;
- ▶ tails, then heads;
- ▶ tails, then tails;

so the probability of obtaining at least one head should be $2/3$, not $3/4$.

Of course, the error in his thinking is that not all of the three outcomes that he cites are equally likely.

Bose's inspiration

If the experiment were modified so that two coins were tossed at the same time, the result remains the same.

But if these two coins were **indistinguishable**, so that the results (h, t) and (t, h) could not be told apart, then d'Alembert's reasoning would be correct.

In the early 1920's, the Indian physicist Satyendra Nath Bose was working on the energy distribution of elementary particles such as photons. Contemporary theory could not explain the experimental data.

In a calculation during a lecture, Bose made a mistake similar to the one described here. He discovered that, based on the mistake, the calculations turned out to correctly describe the data. From this he deduced that photons (and related particles) are indistinguishable – there is in principle no physical way to tell two photons apart.



Satyendra Nath Bose (1894-1974) in Paris, 1925. Photography. Siliconer, August 2000, Vol. 1, 7 File:SatyenBose1925.jpg. (2020, January 27). Wikimedia Commons, the free media repository.

Basic Principles of Counting

Suppose a set A of n objects is given.

- ▶ There are $\frac{n!}{(n-k)!}$ different ways of choosing an ordered tuple of k objects from A .

Such a choice is called a ***permutation of k objects*** from A .

- ▶ There are $\frac{n!}{k!(n-k)!}$ different ways of choosing an unordered set of k objects from A .

Such a choice is called a ***combination of k objects*** from A .

- ▶ There are $\frac{n!}{n_1!n_2!\dots n_k!}$ ways of partitioning A into k disjoint subsets A_1, \dots, A_k whose union is A , where each a_i has n_i elements.

This is called a ***permutation of k indistinguishable objects*** from A .

Binomial Coefficients

We define binomial coefficients by

$$\binom{\alpha}{0} := 1, \quad \text{for } \alpha \in \mathbb{R} \quad (1.1)$$

and, for $n \in \mathbb{N} \setminus \{0\}$ and $\alpha \in \mathbb{R}$,

$$\binom{\alpha}{n} := \frac{\alpha \cdot (\alpha - 1) \cdot (\alpha - 2) \cdots (\alpha - n + 1)}{n!}. \quad (1.2)$$

If $\alpha \in \mathbb{N}$, this may be expressed as the perhaps more familiar

$$\binom{\alpha}{n} = \frac{\alpha!}{(\alpha - n)!n!}.$$

The definition (1.2) also implies that

$$\binom{m}{n} = 0 \quad \text{whenever } n > m \text{ and } m, n \in \mathbb{N}. \quad (1.3)$$

Sample Spaces and Sample Points

We want to translate physical outcomes into mathematical objects, for example:

$$\begin{array}{c} \square \\ \bullet \end{array} \quad \begin{array}{c} \square \\ \bullet\bullet \end{array} \quad \mapsto \quad (1, 3)$$

or

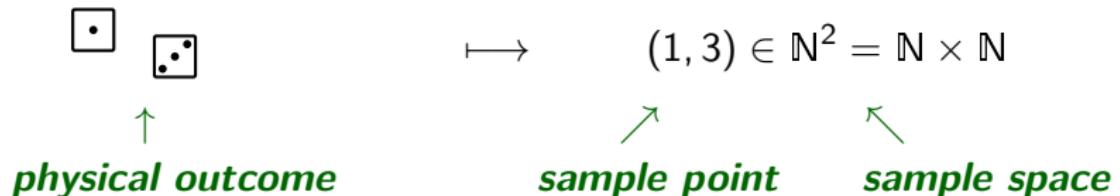

$$\begin{array}{c} \text{person icon} \\ \mapsto \end{array} \quad 173 \text{ cm}$$

The mathematical objects are called **sample points**. They can be numbers, pairs of numbers or any sort of abstract object.

We need to define a **sample space**, often denoted S , large enough to accommodate all sample points.

Events

The sample space can be larger than seems necessary:



An outcome in the sense of Cardano's principle is then interpreted as a subset A of a sample space S and called an **event**.

Two events A_1, A_2 are called **mutually exclusive** if $A_1 \cap A_2 = \emptyset$.

Events

1.4. Example. A six-sided die is rolled four times. The sample space can be taken to be $S = \mathbb{N}^4$ and a sample point is a 4-tuple, for example $(1, 2, 5, 2) \in \mathbb{N}^4$. This sample point would correspond to first rolling a 1, then a 2, next a 5, followed by a 2.

Many tuples, such as $(7, 20, 2, 3) \in S$ do not correspond to any physical outcome of the experiment.

An event might be “rolling at least two fours” in which case this would be a subset $A \subset S$ such that each 4-tuple in A has at least two entries equal to 4. For example, $(1, 3, 4, 4) \in A$ but $(1, 1, 3, 4) \notin A$.

We can then apply counting principles to subsets of sample spaces in order to find the probabilities of events by Cardano’s principle.

Probabilities of Events

1.5. Example. We roll a four-sided die 10 times. What is the probability of obtaining 5 ones, 3 twos, 1 three and 1 four?

There are $4^{10} = 1048576$ possibilities for the 10-tuple of results of the die rolls, corresponding to that many sample points in $S = \mathbb{N}^{10}$ that correspond to physical results. The event A consists of all ordered 10-tuples containing 5 ones, 3 twos, 1 three and 1 four. There are

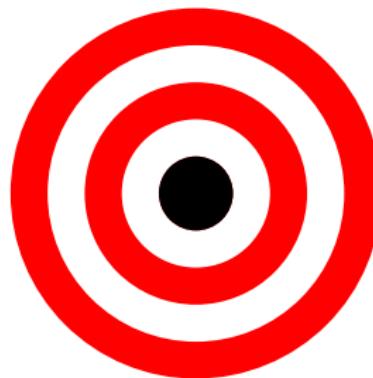
$$\frac{10!}{5!3!1!1!} = 5040$$

possible ways of obtaining 5 ones, 3 twos, 1 three and 1 four, so there are that many elements in A . The probability is

$$\frac{5040}{1048576} \approx 0.00481 \approx 0.5\%.$$

A Dartboard

Consider the dartboard shown below:



The red and white rings have equal width. Suppose a dart is launched and hits a point on the board entirely at random. Hitting a red ring is the event A_1 and hitting a white ring is the event A_2 .

Is $P[A_1] = P[A_2]$? Why or why not? Does the radial symmetry play a role?

An Axiomatic Approach

Clearly, for more complicated situations of randomness that go beyond simple counting, a more formal model of probability is needed. In 1933, the Russian mathematician Kolmogorov introduced an axiomatic approach.

Given a sample space S , we first need to determine the set of permissible events.

If S has a finite number of elements, then we can simply allow any subset of S to be an event. However, if S is very large (for example, if $S = \mathbb{R}$) then a more careful approach is needed.

Not every subset of S may be an allowable event. However, we need to choose “allowable” subsets in a consistent way.



Andrej N. Kolmogorov (1903-1987)

File:Andrej Nikolajewitsch Kolmogorov.jpg.
(2018, December 28). Wikimedia Commons, the free media repository.

A σ -Field of Subsets

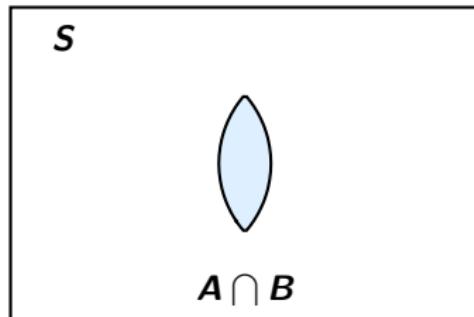
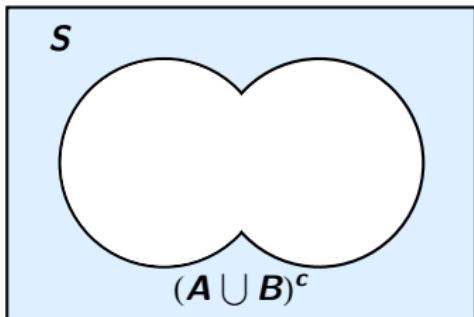
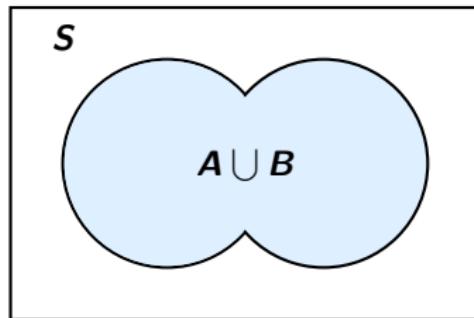
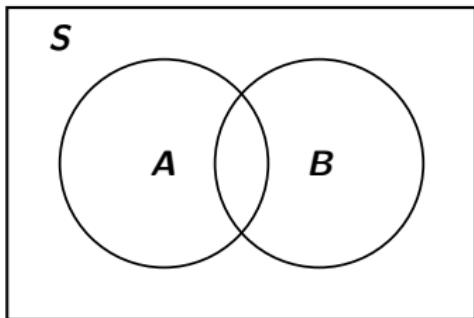
Suppose that a non-empty set S is given. A **σ -field** \mathcal{F} on S is a family of subsets of S such that

- (i) $\emptyset \in \mathcal{F}$;
- (ii) if $A \in \mathcal{F}$, then $S \setminus A \in \mathcal{F}$;
- (iii) if $A_1, A_2, A_3, \dots \in \mathcal{F}$ is a finite or countable sequence of subsets, then the union $\bigcup_k A_k \in \mathcal{F}$.

In probability, we consider families of events that are σ -fields. This is clearly reasonable, since the above properties guarantee that if a subset is an event, then so is the complement and if two subsets are events, then their union must also be an event. Furthermore, the entire set S is an event.

A σ -Field of Subsets

This is illustrated by the diagrams below:



Examples of σ -Fields of Events

- ▶ If S is finite, one can take $\mathcal{F} = \mathcal{P}(S)$ (the power set of S) without problems. This is also the case if S is countable.
- ▶ For any set S , the smallest possible σ -field is $\mathcal{F} = \{\emptyset, S\}$.
- ▶ One of the most important σ -fields in practice is the set $\mathcal{B}(I)$, the set of **Borel sets** on an interval $I \subset \mathbb{R}$. This is the smallest σ -family containing all subintervals of I . (We do not give an explicit definition here.)

Now that we have a sample space and a set of permissible events, we need a probability function that assigns in principle to every event the probability of that event occurring.

Probability Measures and Spaces

Let S be a sample space and \mathcal{F} a σ -field on S . Then a function

$$P: \mathcal{F} \rightarrow [0, 1], \quad A \mapsto P[A],$$

is called a **probability measure** (or **probability function** or just **probability**) on S if

- (i) $P[S] = 1$,
- (ii) For any set of events $\{A_k\} \subset \mathcal{F}$ such that $A_j \cap A_k = \emptyset$ for $j \neq k$,

$$P\left[\bigcup_k A_k\right] = \sum_k P[A_k].$$

The triple (S, \mathcal{F}, P) is called a **probability space**.

Rolling a Die Twice

1.6. Example. Suppose we roll a six-sided die twice. Then we can take the sample space to have 36 elements as follows

$$\begin{aligned} S &= \{(j, k) : j, k = 1, \dots, 6\} \\ &= \{(1, 1), (1, 2), \dots, (6, 5), (6, 6)\}. \end{aligned}$$

We take the σ -field to be the power set $\mathcal{P}(S)$. Following Cardano's approach, we then assign the probability

$$P[\{(i, j)\}] = \frac{1}{36}, \quad \text{for } i, j = 1, 2, 3, 4, 5, 6$$

for each individual sample point. This allows us to define probabilities for an arbitrary event in $\mathcal{P}(S)$.

Rolling a Die Twice

Let A_1 be the event that corresponds to the outcome “the sum of the two die rolls is at most 3” and A_2 correspond to the outcome “the two die rolls give the same number”. Then

$$A_1 = \{(1, 1), (1, 2), (2, 1)\}, \quad A_2 = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}.$$

The probability of these events is calculated as

$$P[A_1] = P[\{(1, 1)\}] + P[\{(1, 2)\}] + P[\{(2, 1)\}] = \frac{3}{36},$$

$$P[A_2] = \frac{6}{36} = \frac{1}{6}.$$

The event “the sum of two die rolls is at most three and both rolls are the same” is given by the set containing those sample points both in A_1 and in A_2 . We calculate its probability to be

$$P[A_1 \cap A_2] = P[\{(1, 1)\}] = \frac{1}{36}.$$

Almost Sure Occurrence

An event $A \in \mathcal{F}$ is said to occur **almost surely** if $P[A] = 1$.

1.7. Example. Suppose we toss a fair coin repeatedly. If it turns heads up, we stop, otherwise we continue to toss. The sample space may be taken to record the tosses as strings of “ t ” (for tails) and “ h ” (for heads), i.e.,

$$S = \{h, th, tth, ttth, \dots\} \cup \{t^\infty\}.$$

where “ t^∞ ” stands for an infinite sequence of tosses yielding tails. This a countable set and we can simply take $\mathcal{F} = \mathcal{P}(S)$.

In order to define a probability function, we set

$$P[\underbrace{t \cdots t}_{n \text{ times}} h] = \frac{1}{2^{n+1}}$$

and $P[\{t^\infty\}] = 0$.

Almost Sure Occurrence

Then the event “Eventually the coin turns up heads.” is given by taking the union of all sample points that include h . We calculate

$$\begin{aligned} P[A] &= P[\{h\}] + P[\{th\}] + P[\{tth\}] + \dots \\ &= \frac{1}{2} + \frac{1}{4} + \frac{1}{8} + \dots \\ &= 1. \end{aligned}$$

We say that “Almost surely, the coin will turn up heads eventually.” However, it is not in principle inconceivable that we toss the coin forever and never see heads turn up. However, the probability of this happening is zero.

Basic Properties of Probabilities of Events

We end by listing some general properties that follow immediately from the definition of a probability space (S, \mathcal{F}, P) :

$$P[S] = 1,$$

$$P[\emptyset] = 0,$$

$$P[S \setminus A] = 1 - P[A],$$

$$P[A_1 \cup A_2] = P[A_1] + P[A_2] - P[A_1 \cap A_2],$$

where $A, A_1, A_2 \in \mathcal{F}$ are any events.

Conditional Probability

Conditional Probability

Given two events A, B in a σ -field \mathcal{F} on a sample space S we can calculate the probability that

- ▶ “event A occurs”,
- ▶ “event A does not occur”,
- ▶ “events A and B occur” and
- ▶ “event A or event B occurs”.

The axioms do not, however, provide us with a way to calculate the probability that

- ▶ “event B occurs if event A has occurred.”

In other words, given information about whether an event A has occurred, we would like to (re-)calculate the probability of B occurring.

Conditional Probability

Let us denote by

$$P[B | A]$$

the **conditional probability** that “ B occurs given that A has occurred”.

2.1. Example. recall from the previous Example 1.6 that in rolling two dice we considered the events

$$A_1 = \{(1, 1), (1, 2), (2, 1)\},$$

$$A_2 = \{(1, 1), (2, 2), (3, 3), (4, 4), (5, 5), (6, 6)\}.$$

What is then $P[A_1 | A_2]$?

If we somehow have the information that the die rolls were equal, we can then conclude that A_1 is only possible if among the six results in A_2 the single result $(1, 1)$ has occurred. We should have

$$P[A_1 | A_2] = \frac{1}{6}.$$

Generalizing the Counting Approach

What we have done is calculate

$$P[A_1 | A_2] = \frac{|A_1 \cap A_2|}{|A_2|}$$

where $|A|$ denotes the number of elements of A . We could re-write this as

$$\begin{aligned} P[A_1 | A_2] &= \frac{|A_1 \cap A_2|/|S|}{|A_2|/|S|} \\ &= \frac{P[A_1 \cap A_2]}{P[A_2]}. \end{aligned}$$

This last expression is independent of the “counting” approach and uses only known probabilistic quantities, so we now define

$$P[B | A] := \frac{P[A \cap B]}{P[A]}$$

whenever $P[A] \neq 0$.

The Two Children Problem



Girl Scout in Uniform File:Girls scout leader in uniform in Trebir, Trebic District.jpg. (2019, September 4).

Suppose that in any birth the probability of the baby being male or female is equal and that the sex is not influenced by that of any siblings.

At a gathering of parents of Girl Scouts, you meet a mother with her daughter. She says “Actually, I also have a second child.”

- (i) What is the probability that the other child is a boy?

The sample space has three elements: $\{(g, g), (b, g), (g, b)\}$, where the first element of the tuple is the gender of the first child, the second element that of the second child. Then we obtain

$$P[\text{other child is a boy}] = \frac{2}{3}.$$

The Two Children Problem

- (ii) She then tells you, "My daughter here is my older child." What is the probability that the other child is a boy?

Now

$$\begin{aligned} & P[\text{other child is a boy} \mid \text{the older child is a girl}] \\ &= \frac{P[\text{other child is a boy and the older child is a girl}]}{P[\text{the older child is a girl}]} \\ &= \frac{1/3}{2/3} = \frac{1}{2}. \end{aligned}$$

The same result would be true if she had said, "My daughter here is my younger child." Even though the daughter must either be younger or older than the second child, knowing which of the two options applies yields a probability of $1/2$ for each sex of the other child. Not knowing the relative age makes it $2/3$ probable that the other child is a boy.

Independence of Events

If one event does not influence another, then we say that the two events are independent. Formally, we say that two events A and B are **independent** if

$$P[A \cap B] = P[A]P[B]. \quad (2.1)$$

Equation (2.1) is equivalent to

$$\begin{aligned} P[A | B] &= P[A] && \text{if } P[B] \neq 0, \\ P[B | A] &= P[B] && \text{if } P[A] \neq 0, \end{aligned}$$

which correspond to the intuitive idea that the probability of A is not affected by B occurring and vice-versa.

The Birthday Problem

2.2. Example. The birthdays (day and month) of a group of people are generally assumed to be independent. Disregarding leap years, any person is assumed to have a $1/365$ chance of being born on a given day. (Do you think that this is a reasonable assumption?) How many people should a group have so that there is a better than even chance of two people in the group having the same birthday?

We consider the complementary problem and start with a single person in the group. If we add a second person, there is a $364/365$ chance of them **not** sharing a birthday. Adding a third person, for no two people to share a birthday, this person must have his birthday on one of the other 363 days of the year, so there is now a

$$\frac{364}{365} \frac{363}{365}$$

chance of no two people in the group sharing a birthday.

The Birthday Problem

Continuing this argument, in a group of $n \geq 2$ people there is a

$$\prod_{k=2}^n \frac{366 - k}{365} = \frac{1}{365^{n-1}} \frac{364!}{(365 - n)!}$$

chance of no two people having the same birthday. It turns out that for $n = 23$ this number is less than 0.5, so the probability of two people having the same birthday is > 0.5 .

This statement has been verified empirically; in a soccer match there are 2×11 players + 1 referee on the pitch. On any given playing day in the Premier Division of the English league, about half the games should feature two participants with the same birthday.

Literature: *Coincidences: The truth is out there*, TEACHING STATISTICS, Vol. 1, No. 1, 1998

Independence vs. Law of Large Numbers

On the one hand, successive flips of a coin are independent - the result of one coin flip should not influence the result of the following coin clips.

On the other hand, experience tells us that if we toss a fair coin many times, it should not come heads up all the time. On average, we expect about one-half of the results to be heads.



Jacob Bernoulli (1654-1705). Painting by Niklaus Bernoulli in 1687. File:Jakob Bernoulli.jpg. (2016, December 29). Wikimedia Commons, the free media repository.

This principle was formulated by Jacob Bernoulli in the early 18th century as the **Law of Large Numbers**:

Probability \longleftrightarrow Proportion of outcomes

Heuristic Version of the Law of Large Numbers

2.3. Heuristic Law of Large Numbers. Let A be a random outcome of an experiment that can be repeated without this outcome influencing subsequent repetitions. Then the probability $P[A]$ of this event occurring may be approximated by

$$P[A] \approx \frac{\text{number of times } A \text{ occurs}}{\text{number of times experiment is performed}}$$

We will give a more precise statement of the law later.

Total Probability

Recall that two events A and B are mutually exclusive if $A \cap B = \emptyset$.

Consider a set of n , pairwise mutually exclusive events A_1, \dots, A_n in a sample space S with the additional properties that $P[A_k] \neq 0$ for all $k = 1, \dots, n$ and $A_1 \cup \dots \cup A_n = S$. Let $B \subset S$ be any event. Then

$$\begin{aligned} P[B] &= P[B \cap S] = P[B \cap (A_1 \cup \dots \cup A_n)] \\ &= P[(B \cap A_1) \cup \dots \cup (B \cap A_n)] \\ &= P[B \cap A_1] + \dots + P[B \cap A_n] \\ &= P[B | A_1] \cdot P[A_1] + \dots + P[B | A_n] \cdot P[A_n] \end{aligned}$$

The expression

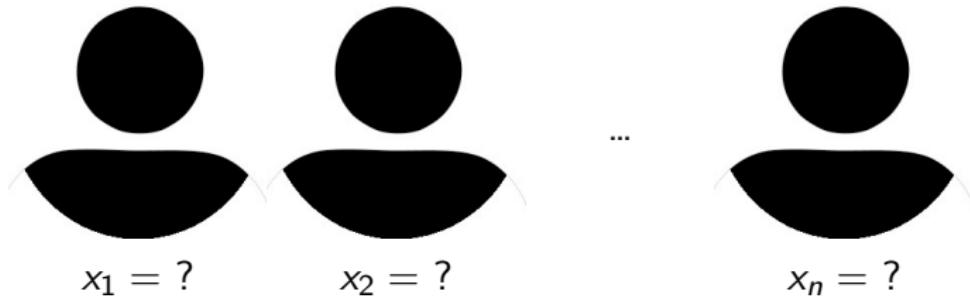
$$P[B] = \sum_{k=1}^n P[B | A_k] \cdot P[A_k]. \quad (2.2)$$

is called the **total probability** formula for $P[B]$.

The Marriage Problem

As an application of the formula for total probability, consider the following **marriage problem**:

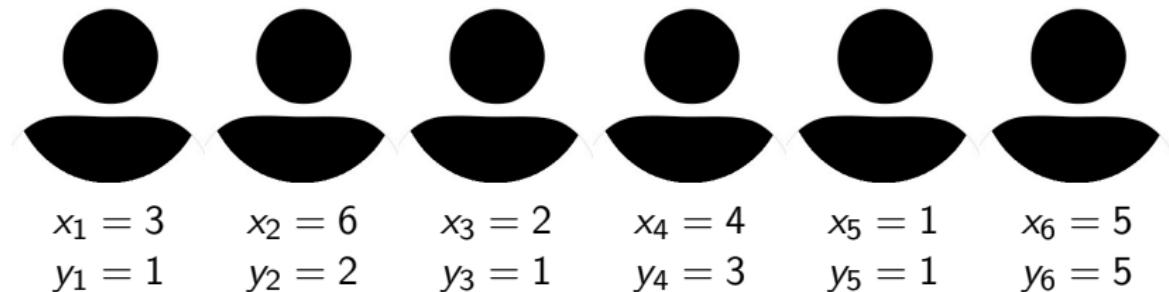
Suppose you are trying to find the “perfect partner.” There are n partners available, and they can be ranked from 1 to n with regard to “suitability”, where the “most suitable” partner has rank 1 and the “least suitable” partner has rank n . We write $x_k \in \{1, \dots, n\}$ for the rank of the k th partner.



The Marriage Problem

You can **evaluate** each partner, but only detect their **relative rank** y_k :
("best so far", "second best so far", etc.).

2.4. Example.



After evaluation: Accept or discard forever.

Goal: Find **most suitable partner** with $x_k = 1$.

Strategy and Outcomes

Optimal strategy: For some $r \geq 1$, evaluate and automatically

reject $r - 1$ potential partners.

Then select the first candidate superior to all the previous ones, if possible.

To summarize:

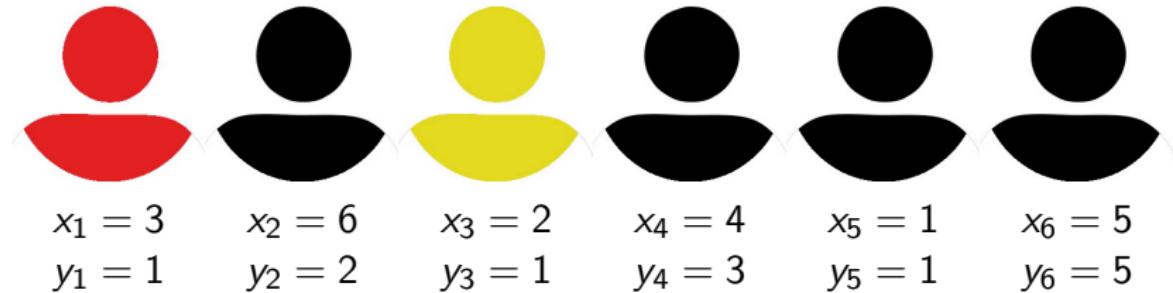
1. Choose $r \geq 1$.
2. Select k with $y_k = 1$ and $k \geq r$, if possible. Discard all others.
3. Otherwise, do not choose anyone.

Possible outcomes:

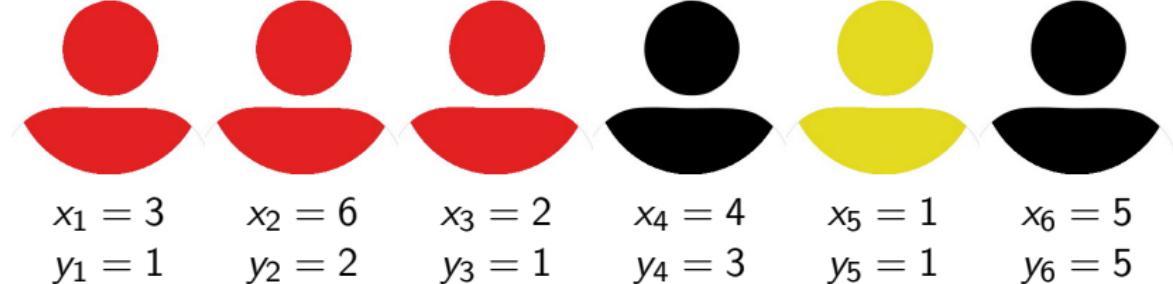
- The most suitable partner ($x_k = 1$) is selected;
- A less suitable partner ($x_k > 1$) is selected;
- No partner is selected.

Examples

$r = 2$:



$r = 4$:



Finding the Optimal Strategy

The sample space can be taken to be

$$S = \{(k, j) : k \text{ is selected and } x_j = 1, \quad k = 0, \dots, n, \quad j = 1, \dots, n\}$$

where $k = 0$ indicates that no person was selected. We say that we “win” if we end up by selecting the most suitable partner, i.e., we select k with $x_k = 1$. We denote this event by

$$W_r = \{(k, k) : k = r, \dots, n\}, \quad r \geq 1.$$

Given $r \geq 1$, the probability of winning is denoted

$$p_r = P[W_r].$$

Problem: **How to choose r so that p_r is maximal?**

The Total Probability Formula

We denote the event that the m th person is the most suitable partner by

$$B_m = \{(k, m) : k = 0, \dots, n\}, \quad m = 1, \dots, n.$$

Note that the B_m are mutually exclusive and that their union is S . Then by the formula for total probability,

$$p_r = P[W_r] = \sum_{m=1}^n P[W_r | B_m]P[B_m].$$

Evaluating the partners in random order,

$$P[B_m] = \frac{1}{n}, \quad m = 1, \dots, n.$$

Probability of Selecting No Partner

No partner will be chosen if and only if the very best candidate was discarded. The probability of this happening is

$$P[\text{selecting no partner}] = \frac{r-1}{n},$$

Of course, in that case we can't win:

$$P[W_r | B_m] = 0 \quad \text{for } m < r$$

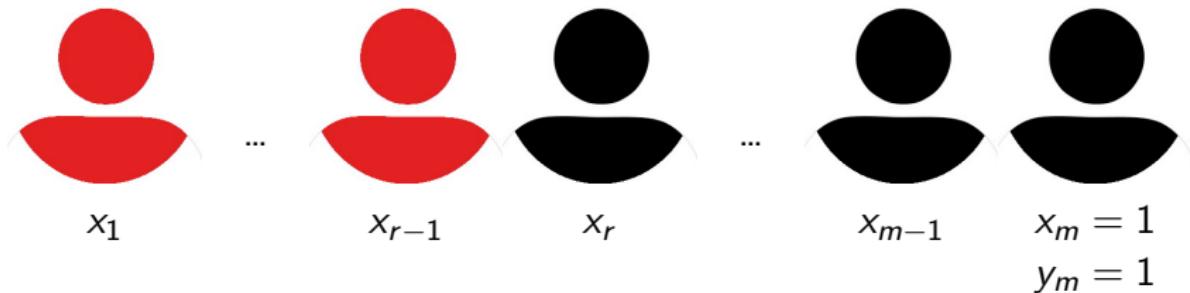
since the most suitable partner will have been discarded.

Therefore, the expression reduces to

$$p_r = \frac{1}{n} \sum_{m=r}^n P[W_r | B_m].$$

The Marriage Problem

Suppose that $x_m = 1$ for $m \geq r$:



We will win if there is no relative rank $y_k = 1$ for $r \leq k < m$.

Hence, the minimum of the finite sequence $(x_1, \dots, x_{r-1}, x_r, \dots, x_{m-1})$ must occur for one of the subscripts $1 \leq k \leq r-1$.

This will happen with probability

$$P[W_r | B_m] = \frac{r-1}{m-1}.$$

The Marriage Problem

We hence have

$$p_r = \frac{r-1}{n} \sum_{m=r}^n \frac{1}{m-1}$$

To find the maximum of this expression, we set $x = r/n$ and use the approximation

$$\sum_{k=1}^n \frac{1}{k} \approx \ln(n) - \gamma$$

where γ is the Euler-Mascheroni constant. Then

$$\begin{aligned} p_r &= \left(x - \frac{1}{n}\right) \left(\sum_{m=2}^n \frac{1}{m-1} - \sum_{m=2}^{r-1} \frac{1}{m-1}\right) \\ &\approx \left(x - \frac{1}{n}\right) (\ln(n-1) - \ln(r-2)) \\ &\approx -x \ln x. \end{aligned}$$

The Marriage Problem

The maximum is now easily found using calculus, yielding $x_{\max} = 1/e$. We hence take $r = \lceil n/e \rceil$, which is about 37% of the partners. Note that p_r has the value $1/e$ at x_{\max} .

The optimal strategy can be summarized as follows: evaluate and reject 37% of the partners, then choose the first partner that is more suitable than any of the preceding partners. This strategy will yield the most suitable partner 37% of the time, lead to no choice (rejection of all partners)

$$\frac{r-1}{n} \approx \frac{r}{n} = x_{\max} = 37\%$$

of the time and lead to an inferior choice 26% of the time.

Bayes's Theorem

From the formula for total probability we immediately obtain one of the most important theorems of elementary probability:

2.5. Bayes's Theorem. Let $A_1, \dots, A_n \subset S$ be a set of pairwise mutually exclusive events whose union is S and who each have non-zero probability of occurring. Let $B \subset S$ be any event such that $P[B] \neq 0$. Then for any A_k , $k = 1, \dots, n$,

$$P[A_k | B] = \frac{P[B \cap A_k]}{P[B]} = \frac{P[B | A_k] \cdot P[A_k]}{\sum_{j=1}^n P[B | A_j] \cdot P[A_j]}.$$

The theorem is due to the English mathematician **Thomas Bayes (1701? - 1761)**. Unfortunately, no clearly authentic image of him survives.

Bayes's Theorem

2.6. Example. Suppose that a rare disease occurs at a rate of 0.1%, i.e., one out of a thousand people have that disease. Suppose a test for the disease is developed that is 99% accurate, i.e., if someone has the disease, the test determines this with 99% accuracy and if someone does not have the disease, the test is negative 99% of the time.

Suppose a patient is tested positive for the disease. What is the probability that she actually has the disease?

We know that

$$P[\text{has disease}] = 0.001,$$

$$P[\text{test positive} \mid \text{has disease}] = 0.99,$$

$$P[\text{test negative} \mid \text{does not have disease}] = 0.99.$$

What we need is $P[\text{has disease} \mid \text{test positive}]$.

Bayes's Theorem

Let us write D for the event “has disease”, $\neg D$ for “does not have disease”, n for “test negative” and p for “test positive”.

By Bayes's Theorem,

$$\begin{aligned} P[\text{has disease} \mid \text{test positive}] &= P[D \mid p] = \frac{P[D \text{ and } p]}{P[p]} \\ &= \frac{P[p \mid D] \cdot P[D]}{P[p \mid D] \cdot P[D] + P[p \mid \neg D] \cdot P[\neg D]} \\ &= \frac{0.99 \cdot 0.001}{0.99 \cdot 0.001 + 0.01 \cdot 0.999} \\ &= 0.0902 \approx 9\%. \end{aligned}$$

Hence, for rare diseases, doctors will always perform a second test on receiving a positive first test. If possible, the second test uses a different principle, so as to be independent of the first test.

The Monty Hall Paradox

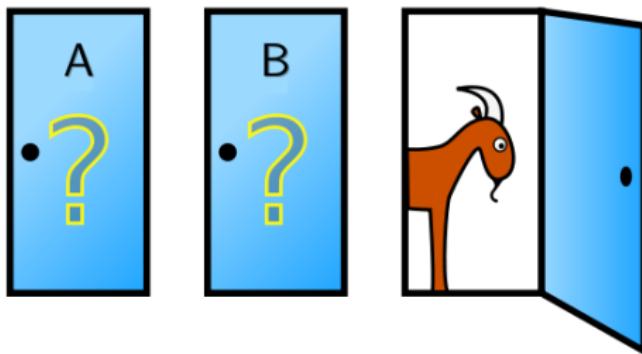
You are participating in a game show to win 10,000,000 RMB. The game master [Monty Hall] presents you with three closed doors. Behind one of the doors is the prize, behind the other two doors there is simply a goat. If you open the correct door, you will receive the money, if you open one of the other two doors you will get a goat.

Before opening any of the three doors, you can announce which door you intend to open. Obviously, at least one of the other two doors does not hide the money. The game master opens this (empty) door. You are then given the option of either

- ▶ sticking with your choice or
- ▶ switching to the other closed door.

What do you do and does it make a difference?

The Monty Hall Paradox



Question. You have chosen door A. Then door C is shown to harbor a goat. To win the money, should you switch to door B?

- A) Yes
- B) No
- C) It doesn't matter.

The Monty Hall Paradox

To many people it seems counter-intuitive, but the best course of action is to **change your choice to the other door**. There will be a $2/3$ probability that the prize is behind the remaining door that you have not chosen!

Why is that? By opening the door, the game master has not given you any information about the door you have chosen (he can always open one of the remaining doors, no matter which door you choose). The probability of this being the correct door was $1/3$ before he opens the other door, and it remains that way after he opens the door.

However, his opening a door **does** give you information on the other two doors, namely, it tells you which of the other two doors does definitely **not** hide the prize. The original $2/3$ probability that one of these doors hides the money is now concentrated on just the one door. Therefore, it is advantageous for you to change your choice.

The Monty Hall Paradox

We can use Bayes's formula to evaluate the probabilities explicitly.

Suppose the doors are denoted A , B and C and denote by the same letter X the event "prize is behind door X " where $X = A, B, C$. Suppose that door A is initially selected and that the host opens door C ; we denote the event "host opens door C " by C^* .

Then, by Bayes's formula,

$$\begin{aligned} P[A \mid C^*] &= \frac{P[C^* \mid A] \cdot P[A]}{P[C^* \mid A] \cdot P[A] + P[C^* \mid C] \cdot P[C] + P[C^* \mid B] \cdot P[B]} \\ &= \frac{\frac{1}{2} \cdot \frac{1}{3}}{\frac{1}{2} \cdot \frac{1}{3} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{3}} = \frac{1}{3}. \end{aligned}$$

Of course, since $P[C \mid C^*] = 0$, this also implies $P[B \mid C^*] = 2/3$.

Discrete Random Variables

Random Variables

Many problems in probability theory revolve around pure numbers rather than arbitrary elements of a sample space (which can be arbitrary objects, such as tuples, or other objects). It is therefore useful to introduce functions that take elements of a sample space and map them into a subset of the real numbers, i.e.,

$$X: S \rightarrow \mathbb{R}.$$

where such a function X is said to be a **random variable**.

The term “random variable” originates from the idea that X has numerical values (“variable”) that are derived from the outcome of a random experiment (“random”).

Random Variables

3.1. Example. Suppose we flip a coin three times. Then the sample space may be given by

$$S = \{(t, t, t), (t, t, h), (t, h, t), (t, h, h), \\ (h, t, t), (h, t, h), (h, h, t), (h, h, h)\}$$

with t denoting “tails” and h denoting “heads”.

We might now define X as follows:

$$X(t, t, t) = 0, \quad X(t, t, h) = 1, \quad X(t, h, t) = 1, \quad X(t, h, h) = 2, \\ X(h, t, t) = 1, \quad X(h, t, h) = 2, \quad X(h, h, t) = 2, \quad X(h, h, h) = 3.$$

Clearly, X denotes the number of heads in three coin flips.

Random Variables

We can now ask what the probability is that X takes on the value 1, which can be found from the probability of each event in the sample space:

$$P[X = 1] = P[\{(t, t, h), (t, h, t), (h, t, t)\}].$$

The notation used on the left is the standard notation for denoting probabilities of random variables. For example, we write

$$P[X = x] = P[A]$$

where $x \in \mathbb{R}$ and $A \subset S$ is the event containing all sample points p such that $X(p) = x$.

More generally, we may write

$$P[a \leq X \leq b]$$

to denote the probability that the values of X lie between a and b .

Random Variables and Probability Density Functions

Hence, the probability that a random variable takes on values in a certain range is in principle determined from the probability space (S, \mathcal{F}, P) .

However, to ensure that this works consistently, a lot of mathematical theory is required if the range of X is (for example) an arbitrary subset of \mathbb{R} .

Therefore, we will make two assumptions:

1) We distinguish between

- ▶ ***discrete random variables***, defined as having a countable range in \mathbb{R}
- ▶ ***continuous random variables***, defined as having range equal to \mathbb{R}

(In principle, a random variable can be of neither of these types, but we will not discuss such cases here.)

2) We assume that a random variable comes with a ***probability density function*** that allows the calculation of probabilities directly, without recourse to the probability space.

Discrete Random Variables

3.2. Definition. Let S be a sample space and Ω a countable subset of \mathbb{R} . A **discrete random variable** is a map

$$X: S \rightarrow \Omega$$

together with a function

$$f_X: \Omega \rightarrow \mathbb{R}$$

having the properties that

- (i) $f_X(x) \geq 0$ for all $x \in \Omega$ and
- (ii) $\sum_{x \in \Omega} f_X(x) = 1$.

The function f_X is called the **probability density function** or **probability distribution** of X .

We often say that a random variable is given by the pair (X, f_X) .

Density for Discrete Random Variables

For discrete random variables, we define the density function f_X in such a way that

$$f_X(x) = P[X = x].$$

In the following slides, we will introduce various concepts based on examples of discrete random variables. We will derive the density function based on the probabilities of the sample space on which the random variables are defined.

Bernoulli Random Variable

Consider an experiment that can result in two possible outcomes, e.g., success or failure, heads or tails, even or odd. Suppose that the probability of success is p , where $0 < p < 1$. Such an experiment is said to be a **Bernoulli trial**.

3.3. Definition. Let S be a sample space and

$$X: S \rightarrow \{0, 1\} \subset \mathbb{R}.$$

Let $0 < p < 1$ and define the density function

$$f_X: \{0, 1\} \rightarrow \mathbb{R}, \quad f_X(x) = \begin{cases} 1 - p & \text{for } x = 0 \\ p & \text{for } x = 1. \end{cases}$$

Then X is said to be a **Bernoulli random variable** or follow a **Bernoulli distribution** with parameter p . We indicate this by writing

$$X \sim \text{Bernoulli}(p)$$

Independent and Identical Trials

More generally, we frequently discuss a sequence of n independent and identical Bernoulli trials. Here,

- ▶ **independent** means that the outcome of one trial does not influence the outcome of the following trials.
- ▶ **identical** means that each trial has the same probability of success.

3.4. Example.

- ▶ If we flip two fair coins, the two trials are independent and identical.
- ▶ If we flip a coin that is fair and another coin that is not fair, the trials are independent but not identical.
- ▶ Suppose a box is filled with 10 red balls and 10 black balls. Twice, we draw a ball out of the box but do not replace it. This is a Bernoulli trial where drawing a red ball counts as a “success”. The probability of success on the first draw is the same as on the second draw (prove this!). Hence the two trials are identical, but they are clearly not independent. (Since the result of the first draw influences the probability of success in the second draw.)

Counting Successes in a Sequence of Trials

Suppose that we perform a sequence of n independent and identical Bernoulli trials. After recording the results, we define X to be the random variable giving the number of successes in n trials.

To determine the density function of X , we need to find the probability of x successes, where $x = 0, 1, \dots, n$. Note that a given sequence of results with x successes occurs with probability

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

since the probability of success is p and the trials are independent and identical. There are $\binom{n}{x}$ ways to place x successes in n trials, hence there are that many sequences with x successes. Since the sequences are mutually exclusive, their probabilities can be added and we find

$$P[x \text{ successes in } n \text{ trials}] = \binom{n}{x} p^x (1 - p)^{n-x}.$$

Binomial Random Variable

3.5. Definition. Let S be a sample space, $n \in \mathbb{N} \setminus \{0\}$, and

$$X: S \rightarrow \Omega = \{0, \dots, n\} \subset \mathbb{R}.$$

Let $0 < p < 1$ and define the density function

$$f_X: \Omega \rightarrow \mathbb{R}, \quad f_X(x) = \binom{n}{x} p^x (1-p)^{n-x}. \quad (3.1)$$

Then X is said to be a ***binomial random variable*** with parameters n and p . We indicate this by writing

$$X \sim B(n, p)$$

Of course, $B(1, p) = \text{Bernoulli}(p)$.

Binomial Random Variable

It is easy to verify that (3.1) is actually a density function: we check that $f_X(x) \geq 0$ for all $x \in \Omega$ and, furthermore,

$$\sum_{x \in \Omega} f_X(x) = \sum_{x=0}^n \binom{n}{x} p^x (1-p)^{n-x} = (p+1-p)^n = 1.$$

We have used the binomial theorem here and this is where the name of the distribution comes from.

Cumulative Distribution Function

In practice, we are also often interested in the **cumulative distribution function** of a random variable, defined as follows,

$$F_X : \mathbb{R} \rightarrow \mathbb{R}, \quad F_X(x) := P[X \leq x]$$

For a discrete random variable

$$F_X(x) = \sum_{y \leq x} f_X(y)$$

and, in particular, in the case of the binomial distribution,

$$F_X(x) = \sum_{y=0}^{\lfloor x \rfloor} \binom{n}{y} p^y (1-p)^{n-y} \quad (3.2)$$

where $\lfloor x \rfloor$ denotes the largest integer not greater than x .



Cumulative Distribution Function

3.6. Example. Suppose a fair coin is tossed 10 times. Then the probability of obtaining not more than three heads is

$$F_X(3) = \sum_{y=0}^3 \binom{10}{y} \frac{1}{2^{10}} = \frac{1 + 10 + 45 + 120}{1024} = \frac{11}{64}$$

There is no simple way of evaluating the sum (3.2), so the values have been tabulated (Table I of Appendix A in the textbook). The Mathematica command for a cumulative distribution function is **CDF**, which for the binomial distribution, however, gives only a representation in terms of a generalized function:

```
CDF[BinomialDistribution[n, p], x]
```

$$\begin{cases} \text{BetaRegularized}[1 - p, n - \text{Floor}[x], 1 + \text{Floor}[x]] & 0 \leq x \leq n \\ 1 & x > n \\ 0 & \text{True} \end{cases}$$

The Geometric Distribution

Let us now look at another example: suppose we perform a sequence of i.i.d. Bernoulli trials which continues until a success is obtained. We then define the **geometric random variable** X to denote the number of trials needed to obtain the first success.

3.7. Example. A fair coin has probability $p = 1/2$ of turning heads up when flipped. The coin is flipped until the first appearance of heads, with the following result: (t, t, t, h) . In this experiment, the geometric random variable X attains the value $X = 4$.

The Geometric Distribution

3.8. Definition. Let S be a sample space and

$$X: S \rightarrow \Omega = \mathbb{N} \setminus \{0\}.$$

Let $0 < p < 1$ and define the density function $f_X: \mathbb{N} \setminus \{0\} \rightarrow \mathbb{R}$ given by

$$f_X(x) = (1 - p)^{x-1} p. \quad (3.3)$$

We say that X is a **geometric random variable** with parameter p and write $X \sim \text{Geom}(p)$.

The cumulative distribution function for a geometrically distributed random variable (X, f_X) with parameter p is given by

$$F(x) = P[X \leq x] = 1 - q^{\lfloor x \rfloor},$$

where $q = 1 - p$ is the probability of failure and $\lfloor x \rfloor$ denotes the greatest integer less than or equal to x .

Probabilities and the Geometric Distribution

In Mathematica, the probability density function f_X is accessed through the command **PDF**, followed by the name of the distribution and the variable of f_X :

```
PDF[GeometricDistribution[p], x]
```

$$\begin{cases} (1-p)^x p & x \geq 0 \\ 0 & \text{True} \end{cases}$$

Note that this differs from (3.3): $x - 1$ is replaced by x . We note: In Mathematica, the geometric distribution gives the **number of failures** before the first success, while our convention is to give the **number of trials** needed for the first success. This is a minor difference and can easily be compensated for, but it illustrates an important point:

When using a computer program, always check that the definitions in the program are the same as the ones you are using!

Probabilities and the Geometric Distribution

The concrete value $f_X(x)$ can be calculated if a given value of x is inserted:

```
PDF[GeometricDistribution[p], 4]
```

$$(1 - p)^4 p$$

Probability can be used to find probabilities such as $P[a \leq x \leq b]$:

```
Probability[1 < x <= 4, x \[Distributed] GeometricDistribution[p]]
```

$$(-1 + p)^2 p (3 - 3 p + p^2)$$

Note that to express equalities as a condition (and not as an assignment of values), Mathematica requires the use of two equals signs:

```
Probability[x == 4, x \[Distributed] GeometricDistribution[p]]
```

$$(-1 + p)^4 p$$



Probabilities and the Geometric Distribution

Question. On August 18, 1913, at the casino in Monte Carlo, a roulette wheel returned black more than 20 times in a row. Find the probability of such an event!

Expectation, Variance and Moments

Averages

Consider the rolling of a fair six-sided die. We are interested in the **average value** of the result. One approach is the following:

Since each result (numbers 1,2,3,4,5,6) occurs with probability 1/6, we take the weighted sum:

$$\frac{1}{6} \cdot 1 + \frac{1}{6} \cdot 2 + \frac{1}{6} \cdot 3 + \frac{1}{6} \cdot 4 + \frac{1}{6} \cdot 5 + \frac{1}{6} \cdot 6 = 3.5.$$

The average result of a die roll is then 3.5, even though this result itself can never occur.

As we shall see later, there are also other ways of thinking of an average (such as the **median** or the **modes** of a distribution).

Expectation

4.1. Definition. Let (X, f_X) be a discrete random variable. Then the **expected value** or **expectation** of X is

$$E[X] := \sum_{x \in \Omega} x \cdot f_X(x).$$

provided that the sum (possibly series, if Ω is infinite) on the right converges absolutely.

We often write μ_X or simply μ for the expectation.

4.2. Example. We will prove later that the expectation for a geometric distribution X is

$$E[X] = \frac{1}{p}$$

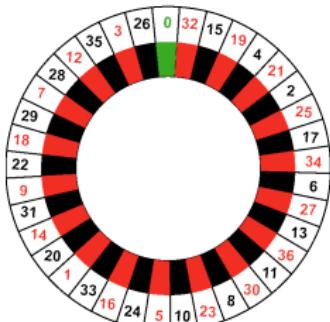
and for a binomial distribution Y the expectation is

$$E[Y] = np.$$

Roulette



Wheel of fortune. Wikimedia Commons.
Wikimedia Foundation. Web. 1 October 2012



Roulette call bets by Darsie. Wikimedia Commons. Wikimedia Foundation. Web. 1 October 2012

In (European) roulette, the player may bet an amount x on a number. If the ball lands on that number, he receives 36 times his initial bet, $36x$; if the ball lands on a different number, he loses his bet. There are 37 numbers on the wheel, so the expected winnings are

$$\begin{aligned} E[W] &= \underbrace{\frac{1}{37} \cdot (-x) + \cdots + \frac{1}{37} \cdot (-x)}_{36 \text{ times}} + \frac{1}{37} \cdot (36 - 1)x \\ &= -\frac{1}{37}x. \end{aligned}$$

A game is said to be **fair** if the expected winnings are zero. The addition of the green zero to the Roulette wheel makes the game into an unfair game for the player and ensures the casino's profit.

American Roulette wheels actually have two zeroes (green "0" and "00")!

St. Petersburg Paradox

Suppose someone offers you the following game: he will flip a fair coin, and if heads come up on the first toss, you receive 2 RMB. If the first toss comes up tails and the second toss comes up heads, he will give you 4 RMB; if only the third toss yields heads, you receive 8 RMB; and so on. Thus, if the first heads comes up on the n th toss, you will receive 2^n RMB.

Question. What is a fair price to pay in order to enter into the game? In other words, what are the expected winnings? How much would you pay to be allowed to play the game?

St. Petersburg Paradox

Calculating the expectation, we see that the probability of the first head coming up on the n th toss is $1/2^n$. Then

$$E[W] = \frac{1}{2} \cdot 2 + \frac{1}{4} \cdot 4 + \frac{1}{8} \cdot 8 + \dots = \infty.$$

The expected value is infinite! (More precisely, the expectation doesn't exist in the sense of Definition 4.1.)

Hence, you should be willing to pay any finite amount (such as 1,000,000 RMB) to participate in the game.

The fact that most people would not pay nearly as much is known as the **St. Petersburg paradox**.

Functions of Random Variables

Given a random variable X , we may consider functions of X . For example,

$$Y := X^2$$

represents the random variable obtained by squaring the values of X . In the case of a discrete random variable, it is not difficult to find the density function f_Y of $Y = \varphi(X)$ where $\varphi: \Omega \rightarrow \mathbb{R}$ represents a suitable function:

$$f_Y(y) = P[Y = y] = P[\varphi(X) = y]$$

In particular, if $y \notin \text{ran } \varphi$, then $P[\varphi(X) = y] = 0$ and hence $f_Y(y) = 0$. Furthermore, since X is discrete,

$$P[\varphi(X) = y] = \sum_{\substack{x \in \Omega \\ \varphi(x)=y}} f_X(x).$$

(Since the outcomes $X = x$ for different values of x are mutually exclusive, their probabilities can be summed.)

Expectation of a Function of a Random Variable

If $X: S \rightarrow \Omega$ is a random variable with density f_X , $\varphi: \Omega \rightarrow \mathbb{R}$ a function and $Y = \varphi(X)$, then

$$\begin{aligned} E[Y] &= \sum_{y \in \mathbb{R}} y \cdot f_Y(y) = \sum_{y \in \mathbb{R}} \sum_{\substack{x \in \Omega \\ \varphi(x)=y}} y \cdot f_X(x) \\ &= \sum_{\substack{(x,y) \in \Omega \times \mathbb{R} \\ y=\varphi(x)}} y \cdot f_X(x) = \sum_{x \in \Omega} \varphi(x) f_X(x). \end{aligned}$$

We have hence proved the following result:

4.3. Lemma. Let (X, f_X) be a discrete random variable and $\varphi: \Omega \rightarrow \mathbb{R}$ some function. Then the expected value of $\varphi \circ X$ is

$$E[\varphi \circ X] = \sum_{x \in \Omega} \varphi(x) \cdot f_X(x).$$

provided that the sum (series) on the right converges absolutely.

Some Properties of the Expectation

By taking $\varphi(x) = c \in \mathbb{R}$ (constant) and $\varphi(x) = c \cdot x$ for some $c \in \mathbb{R}$, we immediately obtain

$$\mathbb{E}[c] = c,$$

$$\mathbb{E}[cX] = c \mathbb{E}[X]$$

for any discrete random variable X .

Given two random variables X and Y their values can be added to yield a new random variable $X + Y$. (This sort of function of multiple random variables will be discussed in more detail in a later section.) For now we give, without proof, the following result:

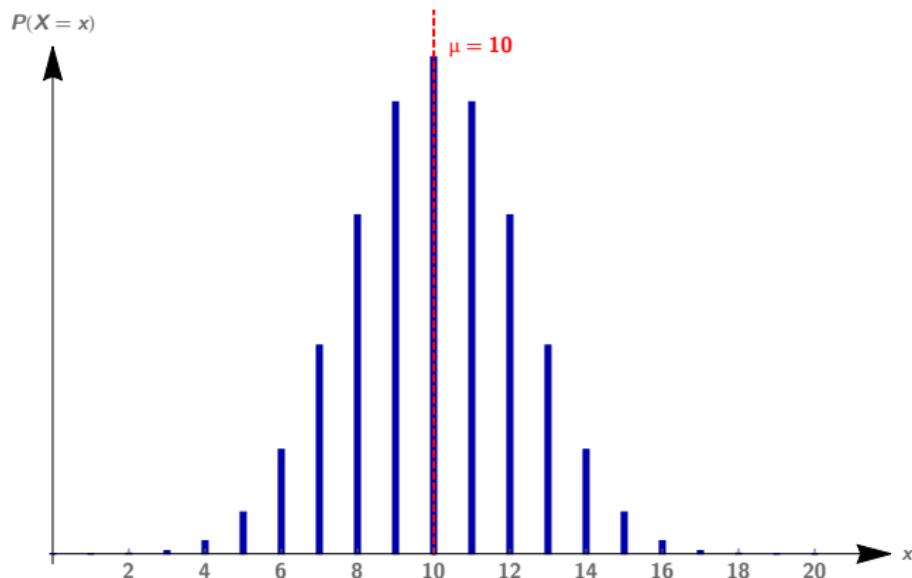
4.4. Theorem. Let X and Y be random variables. Then

$$\mathbb{E}[X + Y] = \mathbb{E}[X] + \mathbb{E}[Y].$$

Location

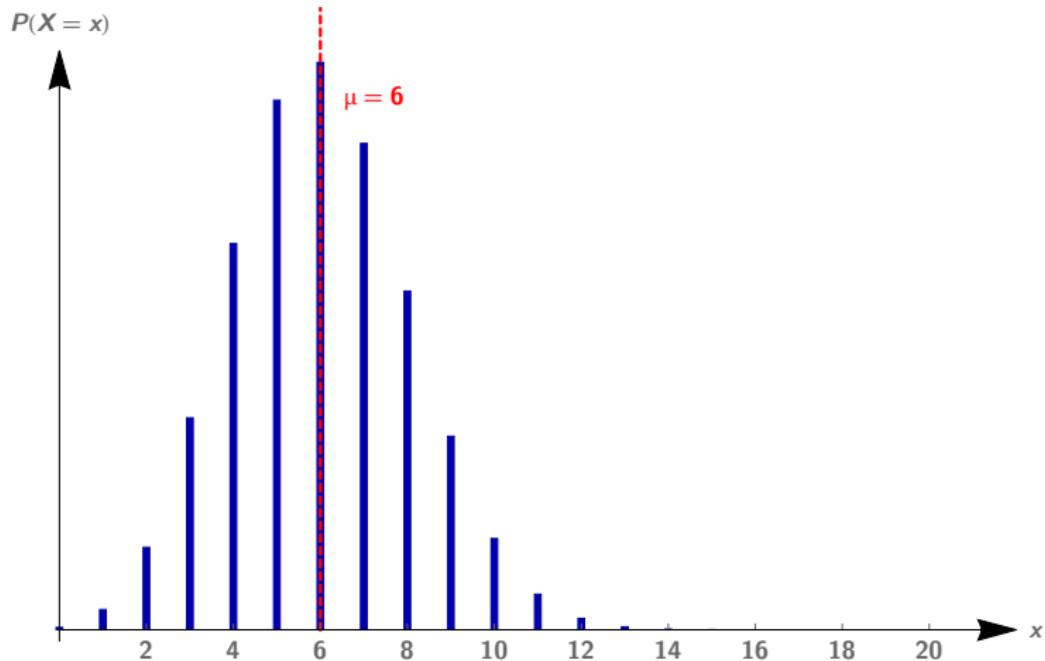
The expectation can be seen as a measure of **location** of a distribution: it indicates where the values of a random variable are concentrated.

The graph below shows the values of the probability density function for a binomial random variable with $n = 20$ and $p = 0.5$:



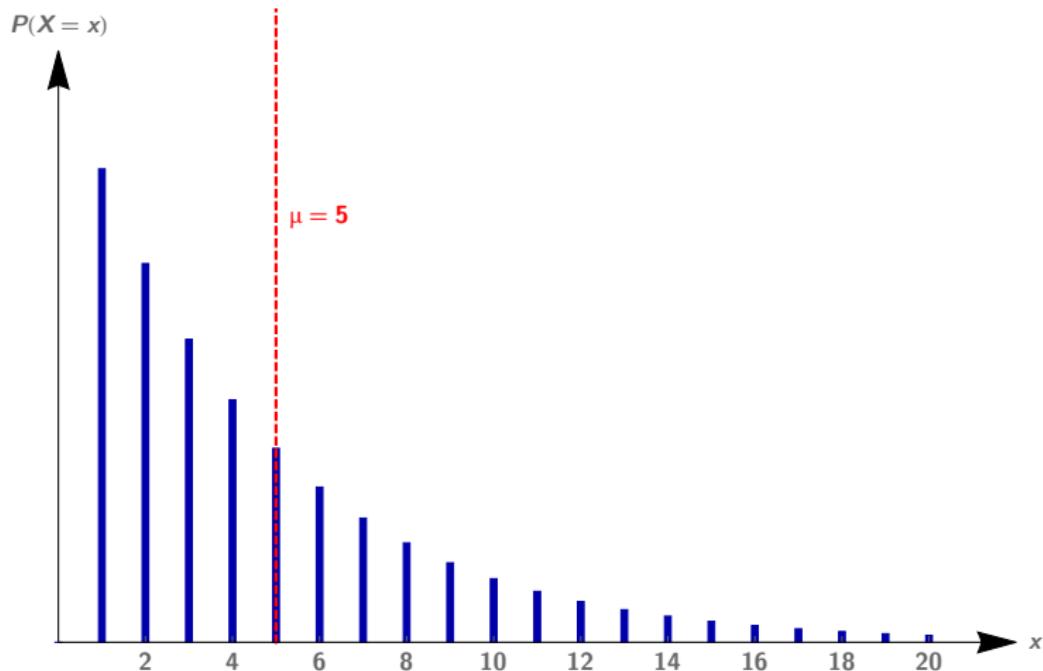
Location

For comparison, here is the graph of the values of the probability density function for a binomial random variable with $n = 20$ and $p = 0.3$:



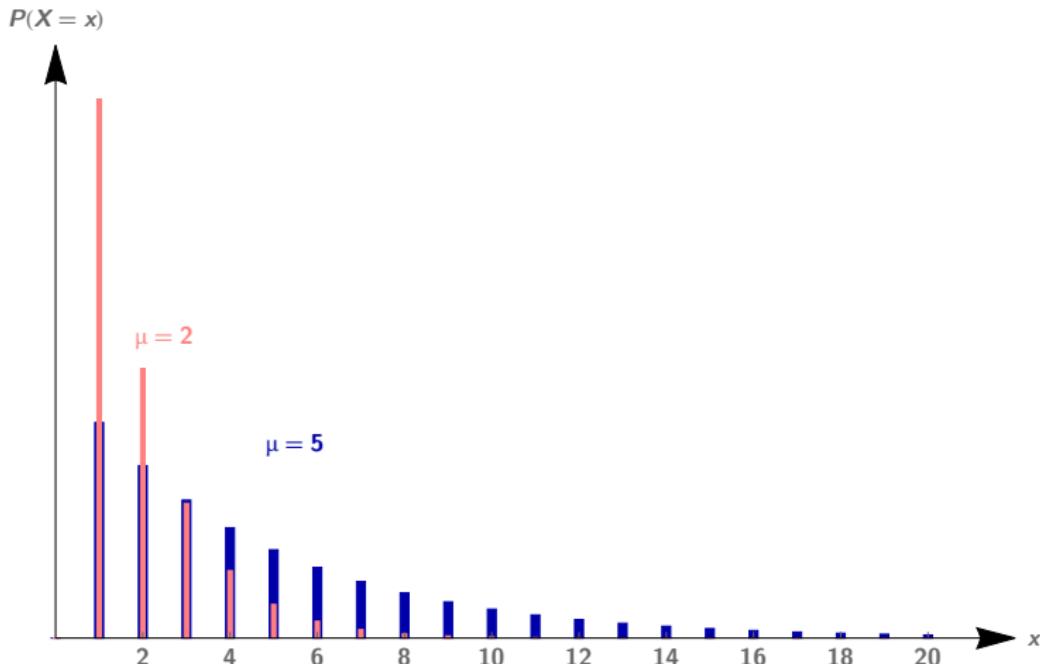
Location

Finally, this is the graph of the values of the probability density function for a geometric random variable with $p = 0.2$:



Dispersion

The **dispersion** measures how much the values of a random variables deviate from their mean. The example below shows two geometric random variables with $p = 0.5$ and $p = 0.2$, respectively.



Variance and Standard Deviation

One possible way to measure the dispersion of a random variable is the ***variance***, which is the

mean square deviation from the mean.

Given X , the deviation from the mean is $X - E[X]$. The mean square deviation is hence

$$\text{Var}[X] := E[(X - E[X])^2],$$

which is defined as long as the right-hand side exists.

The variance is often denoted by σ_X^2 or just σ^2 .

The ***standard deviation*** is defined as

$$\sigma_X = \sqrt{\text{Var}[X]}.$$

Some Properties of the Variance

Using the properties of the mean, we can derive the useful formula

$$\begin{aligned}\text{Var}[X] &= E[(X - E[X])^2] \\ &= E[X^2 - 2E[X] \cdot X + E[X]^2] \\ &= E[X^2] - E[X]^2.\end{aligned}$$

It is then easy to check that for any constant $c \in \mathbb{R}$,

$$\text{Var}[c] = 0, \quad \text{Var}[cX] = c^2 \text{Var}[X]$$

where c by itself is interpreted as a random variable whose values are constant and cX is interpreted in the obvious way.

Standardized Random Variables

It is often useful to “standardize” a random variable by subtracting its mean and dividing by the standard deviation. If X is a given random variable, the standardized variable is hence

$$Y = \frac{X - \mu}{\sigma}.$$

We find that

$$\begin{aligned} E[Y] &= E\left[\frac{1}{\sigma}(X - \mu)\right] = \frac{1}{\sigma} E[(X - \mu)] \\ &= \frac{1}{\sigma}(E[X] - \mu) \\ &= 0 \end{aligned}$$

A similar calculation shows that $\text{Var}[Y] = 1$.

Standardized Random Variables

It is often useful to “standardize” a random variable by subtracting its mean and dividing by the standard deviation. If X is a given random variable, the standardized variable is hence

$$Y = \frac{X - \mu}{\sigma}.$$

We find that

$$\begin{aligned} E[Y] &= E\left[\frac{1}{\sigma}(X - \mu)\right] = \frac{1}{\sigma} E[(X - \mu)] \\ &= \frac{1}{\sigma}(E[X] - \mu) \\ &= 0 \end{aligned}$$

A similar calculation shows that $\text{Var}[Y] = 1$.

Standardized Bernoulli Variables

4.5. Example. Consider a Bernoulli random variable X which takes on values 0 and 1 with probability $p = 1/2$. Then

$$\mathbb{E}[X] = \frac{1}{2} \cdot 0 + \frac{1}{2} \cdot 1 = \frac{1}{2}$$

and

$$\begin{aligned}\text{Var}[X] &= \mathbb{E}[(X - 1/2)^2] = \frac{1}{2}(0 - 1/2)^2 + \frac{1}{2}(1 - 1/2)^2 \\ &= \frac{1}{4}.\end{aligned}$$

Then the standardized random variable is

$$Y = \frac{X - 1/2}{1/2} = 2X - 1.$$

In other words, Y takes on the values 1 and -1 , each with probability $1/2$.

Standardized Random Variables

Question. Given that $Y = (X - \mu)/\sigma$ has expectation zero, what is $E[Y^2]$?

- (a) 0
- (b) 1
- (c) $E[Y^2]$ is not defined.
- (d) $E[Y^2]$ depends on X .

A quick calculation shows that

$$\begin{aligned} E[Y] &= E\left[\frac{1}{\sigma^2}(X - \mu)^2\right] = \frac{1}{\sigma^2} E[(X - \mu)^2] \\ &= \frac{1}{\sigma^2} \text{Var}[X] \\ &= 1. \end{aligned}$$

Ordinary and Central Moments

So far we have encountered the expectation, $E[X]$, and the variance, $\text{Var}[X] = E[X^2] - E[X]^2$. The information contained in these two quantities is basically that of $E[X]$ and $E[X^2]$.

More generally, given a random variable X , the quantities

$$E[X^n], \quad n \in \mathbb{N},$$

are known as the n^{th} (*ordinary*) moments of X .

The quantities

$$E\left[\left(\frac{X - \mu}{\sigma}\right)^n\right], \quad n = 3, 4, 5, \dots,$$

are called the n^{th} (*central*) moments of X .

(Of course, not all moments may exist for a given random variable!)

The Moment-Generating Function

4.6. Definition. Let (X, f_X) be a random variable and such that the sequence of moments $E[X^n]$, $n \in \mathbb{N}$, exists.

If the power series

$$m_X(t) := \sum_{k=0}^{\infty} \frac{E[X^k]}{k!} t^k$$

has radius of convergence $\varepsilon > 0$, the thereby defined function

$$m_X(t): (-\varepsilon, \varepsilon) \rightarrow \mathbb{R}$$

is called the **moment-generating function** for X .

The Moment-Generating Function

4.7. Theorem. Let $\varepsilon > 0$ be given such that $E[e^{tX}]$ exists and has a power series expansion in t that converges for $|t| < \varepsilon$. Then the moment-generating function exists and

$$m_X(t) = E[e^{tX}] \quad \text{for } |t| < \varepsilon.$$

Furthermore,

$$E[X^k] = \left. \frac{d^k m_X(t)}{dt^k} \right|_{t=0}.$$

We can hence calculate the moments of X by differentiating the moment-generating function.

The Moment-Generating Function

The basic idea behind the theorem is to write

$$m_X(t) = \sum_{n=0}^{\infty} \frac{t^n}{n!} E[X^n] = E\left[\sum_{n=0}^{\infty} \frac{t^n X^n}{n!}\right] = E[e^{tX}].$$

The interchange of the infinite series and the expectation is based on property (iii) of Theorem 4.4; however the fact that the series is infinite makes a rigorous justification a little difficult and we omit it here.

Differentiating term-by-term,

$$\frac{d^k m_X(t)}{dt^k} = \sum_{n=0}^{\infty} \frac{d^k}{dt^k} \frac{t^n}{n!} E[X^n] = \sum_{n=k}^{\infty} \frac{t^{n-k}}{(n-k)!} E[X^n].$$

At $t = 0$, only the first term survives, so $\frac{d^k m_X(t)}{dt^k} \Big|_{t=0} = E[X^k]$.

M.G.F. for the Geometric Distribution

It turns out that the moment-generating function is uniquely associated to a given distribution: two random variables will have the same m.g.f. if and only if they have the same probability density function.

We now apply the previous discussion to the geometric distribution:

4.8. Proposition. Let (X, f_X) be a geometrically distributed random variable with parameter p . Then the moment-generating function for X is given by

$$m_X : (-\infty, -\ln q) \rightarrow \mathbb{R}, \quad m_X(t) = \frac{pe^t}{1 - qe^t}$$

where $q = 1 - p$.

M.G.F. for the Geometric Distribution

Proof.

We have $f_X(x) = q^{x-1}p$ for $x \in \mathbb{N} \setminus \{0\}$. Then

$$m_X(t) = E[e^{tX}] = \sum_{x=1}^{\infty} e^{tx} q^{x-1} p = \frac{p}{q} \sum_{x=1}^{\infty} (qe^t)^x$$

This is a geometric series which converges for $|qe^t| = qe^t < 1$, i.e., for $t < -\ln q$. For such t , the limit is given by

$$\begin{aligned} m_X(t) &= \frac{p}{q} \sum_{x=1}^{\infty} (qe^t)^x = \frac{p}{q} \left(\sum_{x=0}^{\infty} (qe^t)^x - 1 \right) = \frac{p}{q} \left(\frac{1}{1 - qe^t} - 1 \right) \\ &= \frac{p}{q} \frac{qe^t}{1 - qe^t} = \frac{pe^t}{1 - qe^t}. \end{aligned}$$



Expectation and Variance for the Geometric Distribution

4.9. Lemma. Let (X, f_X) be a geometrically distributed random variable with parameter p . Then the expectation value and variance are given by

$$E[X] = \frac{1}{p} \quad \text{and} \quad \text{Var}[X] = \frac{q}{p^2}$$

where $q = 1 - p$.

Proof.

We use the moment-generating function to calculate the expectation value:

$$\begin{aligned} E[X] &= \frac{d}{dt} \Big|_{t=0} m_X(t) = \frac{d}{dt} \Big|_{t=0} \frac{p}{e^{-t} - q} \\ &= \frac{pe^t(1 - qe^t) + pq}{(e^{-t} - q)^2} \Big|_{t=0} = \frac{p}{(1 - q)^2} = \frac{1}{p}. \end{aligned}$$

The proof for the variance is similar and is left to the reader. □



Expectation and Variance for the Binomial Distribution

4.10. Theorem. Let (X, f_X) be a binomial random variable with parameters n and p .

- (i) The moment generating function of X is given by

$$m_X: \mathbb{R} \rightarrow \mathbb{R}, \quad m_X(t) = (q + pe^t)^n, \quad q = 1 - p.$$

(ii) $E[X] = np.$

(iii) $\text{Var}[X] = npq.$

The proof of this theorem is left as an exercise.

The Mathematica commands for the expectation and variance are:

```
Mean[BinomialDistribution[n, p]]
```

$n p$

```
Variance[BinomialDistribution[n, p]]
```

$n (1 - p) p$

The Pascal, Negative Binomial and Poisson Distributions

Generalizing the Geometric Distribution

Consider a sequence of independent, identical Bernoulli trials with probability $0 < p < 1$ of success.

Question: How many trials are necessary to obtain $r > 0$ successes, where r is a fixed parameter?



(The situation described by the geometric distribution corresponds to the case $r = 1$ here.)

We calculate the probability that $x \geq r$ trials are needed to obtain r successes.

Counting Trials for r Successes

Main idea: If the r^{th} success is obtained in the x^{th} trial, then there must have been ***exactly $r - 1$ successes in the previous $x - 1$ trials.***

$$P[\text{exactly } r - 1 \text{ successes in } x - 1 \text{ trials}] = \binom{x-1}{r-1} p^{r-1} (1-p)^{x-r}.$$

Now with probability p the x^{th} trial will be a success, so

$$P[\text{obtain } r^{\text{th}} \text{ success in the } x^{\text{th}} \text{ trial}] = \binom{x-1}{r-1} p^r (1-p)^{x-r}.$$

The Pascal Distribution

5.1. Definition. Let $r \in \mathbb{N} \setminus \{0\}$. A random variable (X, f_X) with

$$\begin{aligned} X: S &\rightarrow \Omega = \mathbb{N} \setminus \{0, 1, \dots, r-1\} \\ &= \{r, r+1, r+2, \dots\} \end{aligned}$$

and distribution function $f_X: \Omega \rightarrow \mathbb{R}$ given by

$$f_X(x) = \binom{x-1}{r-1} p^r (1-p)^{x-r}, \quad 0 < p < 1,$$

is said to follow a **Pascal distribution** with parameters p and r .



Blaise Pascal (1623-1672): Anonym. ca. 1690. Painting. Palais de Versailles. Paris. File:Blaise Pascal Versailles.JPG. (2020, February 12). Wikimedia Commons, the free media repository.

The Negative Binomial Distribution

Instead of counting the number of trials needed to obtain r successes, we may count ***the number of failures obtained before r successes***:

5.2. Definition. Let $r \in \mathbb{N} \setminus \{0\}$. A random variable (X, f_X) with

$$X: S \rightarrow \Omega = \mathbb{N}$$

and distribution function $f_X: \Omega \rightarrow \mathbb{R}$ given by

$$f_X(x) = \binom{x+r-1}{r-1} p^r (1-p)^x, \quad 0 < p < 1,$$

is said to follow a ***negative binomial distribution*** with parameters p and r .

The Negative Binomial Distribution

The term “negative binomial” comes from the fact that

$$\begin{aligned}\binom{-r}{x} &= \frac{(-r) \cdot (-r - 1) \cdots (-r - x + 1)}{x!} \\ &= \frac{r \cdot (r + 1) \cdots (r + x - 1)}{x!} (-1)^x \\ &= (-1)^x \frac{(r + x - 1)!}{x!(r - 1)!} = (-1)^x \binom{r - 1 + x}{r - 1}\end{aligned}$$

so that the density of the negative binomial distribution may be expressed as

$$f_X(x) = \binom{-r}{x} (-1)^x p^r (1 - p)^x$$

We now return to the Pascal distribution.



The M.G.F. for the Pascal Distribution

5.3. Theorem. Let (X, f_X) be a Pascal random variable with parameters p and r .

(i) The moment generating function of X is given by

$$m_X : (-\infty, -\ln q) \rightarrow \mathbb{R}, \quad m_X(t) = \frac{(pe^t)^r}{(1 - qe^t)^r}, \quad q = 1 - p.$$

(ii) $E[X] = r/p$.

(iii) $\text{Var}[X] = rq/p^2$.

Using Mathematica:

```
MomentGeneratingFunction[PascalDistribution[r, p], t]
```

$$\left(\frac{e^t p}{1 - e^t (1 - p)} \right)^r$$

The M.G.F. for the Pascal Distribution

Proof.

We derive the moment-generating function only. It is given by

$$\begin{aligned}m_X(t) &= E[e^{Xt}] = \sum_{x=r}^{\infty} e^{tx} \binom{x-1}{r-1} p^r (1-p)^{x-r} \\&= \sum_{x=0}^{\infty} e^{t(r+x)} \binom{r+x-1}{r-1} p^r (1-p)^x \\&= p^r e^{tr} \sum_{x=0}^{\infty} \binom{-r}{x} [-e^t(1-p)]^x\end{aligned}$$

Recall the **binomial series**

$$(1-y)^{-r} = \sum_{x=0}^{\infty} \binom{-r}{x} (-y)^x \quad \text{for } |y| < 1.$$

The M.G.F. for the Pascal Distribution

Proof (continued).

It follows that, as long as $e^t(1 - p) < 1$,

$$\begin{aligned}m_X(t) &= p^r e^{tr} \sum_{x=0}^{\infty} \binom{-r}{x} [-e^t(1-p)]^x \\&= p^r e^{tr} (1 - (1-p)e^t)^{-r} = \frac{(pe^t)^r}{(1-qe^t)^r}\end{aligned}$$

with $q = 1 - p$.

□

5.4. Remark. A random variable following the Pascal distribution with parameters r and p is the sum of r independent geometric random variables with parameter p .

Counting Successes in a Continuous Environment

Binomial distribution: counts successes in n trials.

Now: count successes in a continuous interval $[a, b] \subset \mathbb{R}$.

Examples:

- ▶ number of earthquakes in a century;
- ▶ number of child births in a day;
- ▶ number of bacteria in a unit volume of water.

We will talk about **arrivals in a time interval** $[0, t]$ for some $t > 0$. The number of arrivals will be denoted by X_t .

Assumptions:

- (i) **Independence:** If the intervals $T_1, T_2 \subset [0, t]$ do not overlap (except perhaps at one point), then the numbers of arrivals in these intervals are independent of each other.
- (ii) **Constant rate of arrivals.**

Rate of Arrivals (Heuristic Postulates)

Assumption: There exists a number $\lambda > 0$ (arrival rate) such that for any small time interval of size Δt the following postulates are satisfied:

- (i) The probability that exactly one arrival will occur in an interval of width Δt is approximately $\lambda \cdot \Delta t$.
- (ii) The probability that exactly zero arrivals will occur in the interval is approximately $1 - \lambda \cdot \Delta t$.
- (iii) The probability that two or more arrivals occur in the interval is approximately zero (very small).

Wanted: a more precise (mathematical) expression of these principles.

Rate of Arrivals (Heuristic Postulates)

5.5. Example. If a hospital ward experiences, on average, about 12 child births per day, spread completely randomly throughout 24 hours, then in any given 10-minute period

- (i) The probability that exactly one child birth will occur is approximately

$$\lambda \cdot \Delta t = \frac{12}{24 \text{ hours}} \cdot \frac{1}{6} \text{ hours} = \frac{1}{12}.$$

- (ii) The probability that exactly zero births will occur is approximately

$$1 - \lambda \cdot \Delta t = \frac{11}{12}$$

- (iii) The probability that two or more births occur is approximately zero (very small).

“Little-o” Notation

We denote by $o(t)$ any function f such that

$$\lim_{t \rightarrow 0} \frac{f(t)}{t} = 0.$$

Hence $o(t)$ does not denote a particular function, rather a class of functions. For example,

- ▶ $t^2 = o(t)$,
- ▶ $(1+t)^2 = 1 + 2t + o(t)$,
- ▶ $\sin t = t + o(t)$.

In particular,

- ▶ $o(t) + o(t) = o(t)$,
- ▶ $t^n \cdot o(t) = o(t)$ for all $n \in \mathbb{N}$,
- ▶ $o(t) \cdot o(t) = o(t)$.

Rate of Arrivals (Precise Postulates)

- (i) The probability that ***exactly one arrival*** will occur in an interval of width Δt is

$$\lambda \cdot \Delta t + o(\Delta t).$$

- (ii) The probability that ***exactly zero arrivals*** will occur in the interval is

$$1 - \lambda \cdot \Delta t + o(\Delta t).$$

- (iii) The probability that ***two or more arrivals*** occur in the interval is

$$o(\Delta t).$$

We denote by X_t the number of arrivals in the interval $[0, t]$ and write

$$P[X_t = x] =: p_x(t) \quad \text{with } x = 0, 1, 2, 3, \dots$$

Probability of Zero Arrivals

Consider the time interval

$$[0, t + \Delta t] = [0, t] \cup [t, \Delta t]$$

Due to independence of non-overlapping time intervals,

$$\begin{aligned} p_0(t + \Delta t) &= P[0 \text{ arrivals in } [0, t + \Delta t]] \\ &= P[0 \text{ arrivals in } [0, t]] \cdot P[0 \text{ arrivals in } [t, t + \Delta t]] \\ &= p_0(t) \cdot (1 - \lambda \Delta t + o(\Delta t)) \end{aligned}$$

It follows that

$$-\lambda p_0(t) = \frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} + \frac{o(\Delta t)}{\Delta t}.$$

Probability of Zero Arrivals

We can take the limit as $\Delta t \rightarrow 0$ on both sides. Then we have

$$-\lambda p_0(t) = \lim_{\Delta t \rightarrow 0} \frac{p_0(t + \Delta t) - p_0(t)}{\Delta t} = p'_0(t).$$

This is a linear, homogeneous ordinary differential equation for p_0 .

Question. What is a suitable initial value for this ODE?

- (a) 0.
- (b) 1.
- (c) Some other value.

Probability of Several Arrivals

Now let $x > 0$. Then

$$\begin{aligned} p_x(t + \Delta t) &= P[x \text{ arrivals in } [0, t + \Delta t]] \\ &= \sum_{y=0}^x P[x - y \text{ arrivals in } [0, t]] \cdot P[y \text{ arrivals in } [t, t + \Delta t]] \\ &= p_x(t) \cdot (1 - \lambda\Delta t + o(\Delta t)) + p_{x-1}(t) \cdot (\lambda\Delta t + o(\Delta t)) \\ &\quad + p_{x-2}(t) \cdot o(\Delta t) + \cdots + p_0(t) \cdot o(\Delta t) \\ &= \lambda\Delta t p_{x-1}(t) + (1 - \lambda\Delta t)p_x(t) + o(\Delta t) \end{aligned}$$

so that

$$\lambda p_{x-1}(t) - \lambda p_x(t) = \frac{p_x(t + \Delta t) - p_x(t)}{\Delta t} + \frac{o(\Delta t)}{\Delta t}.$$

Probability of Several Arrivals

Taking the limit as $\Delta t \rightarrow 0$, we obtain

$$p'_x(t) = \lambda p_{x-1}(t) - \lambda p_x(t).$$

Together with

$$p'_0 = -\lambda p_0$$

and suitable initial conditions we have a system of differential equations that can be solved inductively to determine p_0, p_1, p_2, \dots .

The solution to these equations is

$$p_x(t) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}.$$

We often define $k := \lambda t$ ("rate times interval").

The Poisson Distribution



Siméon Poisson (1781-1840). Delpech, François Séraphin before 1840. Lithograph. File:Siméon Poisson.jpg. (2019, August 13). Wikimedia Commons, the free media repository.

5.6. Definition. Let $k \in \mathbb{R}$. A random variable (X, f_X) with

$$X: S \rightarrow \mathbb{N}$$

and density function $f_X: \mathbb{N} \rightarrow \mathbb{R}$ given by

$$f_X(x) = \frac{k^x e^{-k}}{x!}$$

is said to follow a **Poisson distribution** with parameter k .

The Poisson distribution describes the occurrence of events that occur at a **constant rate** in a **continuous environment**.

M.G.F. and C.D.F. of the Poisson Distribution

5.7. Theorem. Let (X, f_X) be a Poisson distributed random variable with parameter k .

- (i) The moment generating function of X is given by

$$m_X: \mathbb{R} \rightarrow \mathbb{R}, \quad m_X(t) = e^{k(e^t - 1)}.$$

- (ii) $E[X] = k$.
(iii) $\text{Var}[X] = k$.

The cumulative distribution function

$$F(x) = P[X \leq x] = \sum_{y=0}^{\lfloor x \rfloor} \frac{e^{-k} k^y}{y!}$$

is found in Table II of Appendix A of the text book.

The Poisson Distribution

5.8. Example. A healthy individual may have an average white blood cell count of as low as $4500/\text{mm}^3$ of blood. To detect a white-cell deficiency, a 0.001 mm^3 drop of blood is taken and the number X of white blood cells is found.

If at most one is found, is there evidence of a white-cell deficiency?

Here the volume of blood (in mm^3) takes the role of the continuous variable and each observed white cell counts as an “arrival.”

The number of arrivals per unit volume is $\lambda = 4500$, the volume under consideration is $s = 0.001$. Hence we have a Poisson-distributed random variable with parameter

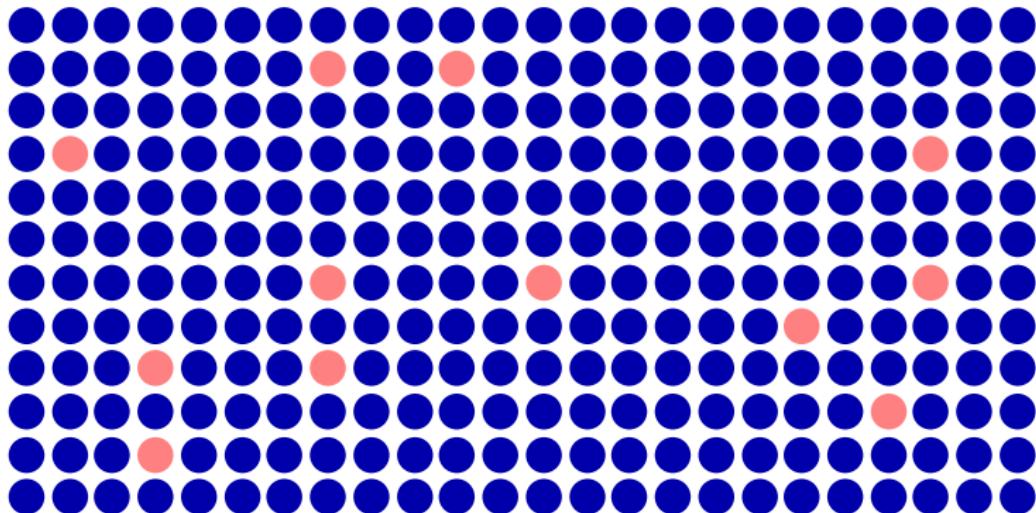
$$k = \lambda s = 4.5.$$

The expected value is $E[X] = k = 4.5$. Furthermore,

$$P[X \leq 1] = \sum_{x=0}^1 \frac{e^{-4.5} 4.5^x}{x!} = 0.061.$$

Approximating the Binomial Distribution

Suppose a binomial random variable is given with **large n** . Then we can approximate the density function using that of a Poisson distribution:



Within many trials (represented by each disk) the successes (orange disks) occur as within a continuum of trials.

Approximating the Binomial Distribution

Mathematically, this is actually a limit statement: If $n \rightarrow \infty$ while $n \cdot p =: \lambda$ remains constant,

$$\binom{n}{m} p^m (1-p)^{n-m} \xrightarrow[n \cdot p = k]{n \rightarrow \infty} \frac{k^m}{m!} e^{-k}$$

Therefore, we can approximate the binomial distribution by a Poisson distribution with parameter

$$k = pn$$

if n is large.

In general, one does this if $p < 0.1$. The smaller p and the larger n are, the better the approximation.

Approximating the Binomial Distribution



5.9. Example. A typical aircraft wing has 40, 000 rivets. Suppose that the probability of a given rivet being defective is 0.001. What is the probability that not more than fifty rivets are defective?

The actual probability is

$$\begin{aligned} P[X \leq 50] &= \sum_{x=0}^{50} \binom{40\,000}{x} (0.001)^x (0.999)^{40\,000-x} \\ &= 0.94746. \end{aligned}$$

Using the Poisson approximation, $k = 40\,000 \cdot 0.001 = 40$ and

$$P[X \leq 50] \approx \sum_{x=0}^{50} e^{-40} \frac{40^x}{x!} = 0.94737.$$

Continuous Random Variables

Continuous Random Variables

6.1. Definition. Let S be a sample space. A **continuous random variable** is a map $X: S \rightarrow \mathbb{R}$ together with a function $f_X: \mathbb{R} \rightarrow \mathbb{R}$ with the properties that

(i) $f_X(x) \geq 0$ for all $x \in \mathbb{R}$ and

(ii) $\int_{-\infty}^{\infty} f_X(x) dx = 1.$

The integral of f_X is interpreted as the probability that X assumes values x in a given range, i.e.,

$$P[a \leq X \leq b] = \int_a^b f_X(x) dx$$

The function f_X is called the **probability density function** (or just density) of the random variable X .

The Probability Density Function

Notice that by the above definition,

$$P[X = x] = \int_x^x f_X(y) dy = 0,$$

i.e., the probability that X assumes any specific value is zero. We see that f_X no longer represents a probability, but is truly a *density*.

Cumulative Distribution

6.2. Definition. Let (X, f_X) be a continuous random variable. The cumulative distribution function for X is defined by $F_X: \mathbb{R} \rightarrow \mathbb{R}$,

$$F_X(x) := P[X \leq x] = \int_{-\infty}^x f_X(y) dy$$

Notice that by the fundamental theorem of calculus we can easily obtain the density f_X from F_X :

$$f_X(x) = F'_X(x).$$

Expectation and Variance

We can define the expectation of a continuous random variable X analogously to that of discrete variables:

$$E[X] := \int_{\mathbb{R}} x \cdot f_X(x) dx$$

It is possible to prove (using some technical arguments in measure theory) that for any “reasonable” function $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ we have

$$E[\varphi \circ X] = \int_{-\infty}^{\infty} \varphi(x) \cdot f_X(x) dx,$$

similarly to the discrete case. As before,

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

and all the previously established properties of the expectation and variance continue to hold in the continuous case.

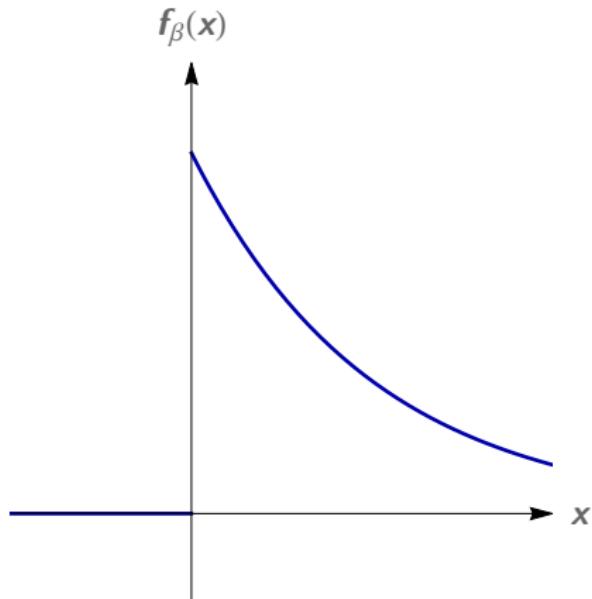
The Exponential Distribution

6.3. Definition. Let $\beta \in \mathbb{R}$, $\beta > 0$.

A continuous random variable (X, f_β) with density

$$f_\beta(x) = \begin{cases} \beta e^{-\beta x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

is said to follow an **exponential distribution** with parameter β .



It is easy to verify that $f_\beta(x) \geq 0$ for all $x \in \mathbb{R}$ and

$$\int_{-\infty}^{\infty} f_\beta(x) dx = 1.$$

Expectation and Variance

Through integration by parts, we find the expectation and variance:

$$\begin{aligned} E[X] &= \int_{-\infty}^{\infty} x f_{\beta}(x) dx = \int_0^{\infty} \beta x e^{-\beta x} dx \\ &= -xe^{-\beta x} \Big|_0^{\infty} + \int_0^{\infty} e^{-\beta x} dx = \frac{1}{\beta}. \end{aligned}$$

The second moment is

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f_{\beta}(x) dx = \int_0^{\infty} \beta x^2 e^{-\beta x} dx \\ &= -x^2 e^{-\beta x} \Big|_0^{\infty} + 2 \int_0^{\infty} x e^{-\beta x} dx = \frac{2}{\beta^2} \end{aligned}$$

and therefore,

$$\text{Var}[X] = E[X^2] - E[X]^2 = \frac{1}{\beta^2}.$$

The Moment-Generating Function

We now see that

$$m_X(t) = E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_X(x) dx$$

so the moment-generating function of a continuous random variable is (up to a sign) the **bilateral Laplace transform** of its density.

For the exponential distribution we have

$$\begin{aligned} m_X(t) &= E[e^{tX}] = \int_{-\infty}^{\infty} e^{tx} f_{\beta}(x) dx \\ &= \int_0^{\infty} \beta e^{-(\beta-t)x} dx \\ &= \frac{\beta}{(\beta - t)} \int_0^{\infty} e^{-y} dy \\ &= (1 - t/\beta)^{-1}. \end{aligned}$$

Connection to the Poisson Distribution

The exponential distribution has a close relationship with the Poisson distribution.

Recall that for Poisson-distributed events (arrivals) the probability of x arrivals in the time interval $[0, t]$ is given by

$$p_x(t) = \frac{(\lambda t)^x}{x!} e^{-\lambda t}, \quad x \in \mathbb{N}.$$

Then $p_0(t)$ is the probability of no arrivals in $[0, t]$. This can also be interpreted as the probability that the first arrival occurs at a time greater than t .

Denote by T the time of the first arrival (it is a continuous random variable). Then

$$P[T > t] = p_0(t) = e^{-\lambda t}, \quad t \geq 0.$$

and $P[T > t] = 1$ for $t < 0$.

Connection to the Poisson Distribution

Hence, if we denote by F_T the cumulative distribution of the density of T , we have

$$F_T(t) = P[T \leq t] = 1 - e^{-\lambda t}, \quad t \geq 0,$$

and $F_T(t) = 0$ for $t < 0$. Since $f_T(t) = F'_T(t)$, the density is

$$f_T(t) = \lambda e^{-\lambda t}, \quad t \geq 0.$$

and $f_T(t) = 0$ for $t < 0$.

Thus the time between successive arrivals of a Poisson-distributed random variable is exponentially distributed with parameter $\beta = \lambda$.

Connection to the Poisson Distribution

6.4. Example. An electronic component is known to have a useful life represented by an exponential density with failure rate of $\lambda = 10^{-5}$ failures per hour, i.e., $\beta = 10^{-5}$. The mean time to failure, $E[X]$, is thus $1/\beta = 10^5$ hours.

Suppose we wanted to determine the fraction of such components that would fail before the mean or expected life:

$$P[T \leq 1/\beta] = \int_0^{1/\beta} \beta e^{-\beta x} dx = 1 - e^{-1} = 0.63212.$$

That is, 63.2% of the components will fail before the mean life time.

Observe that this result does not depend on the value of β .

Location of Continuous Distributions

This is a good opportunity to discuss the **location** of a random variable (X, f_X) . The location is supposed to give the “center” of the distribution. There are three main ways of doing this:

- (i) The **median** M_X , defined by $P[X \leq M_X] = 0.5$. In the context of Example 6.4, this is the time where half of the components will have failed.
- (ii) The **mean** $E[X]$.
- (iii) The **mode** x_0 , which is the location of the maximum of f_X (if there is a unique maximum location). In the context of Example 6.4, the mode gives the time with the greatest failure density, i.e., the time around which failure is most likely. For the exponential distribution, $x_0 = 0$.

Depending on the application, any of these three measures may be referred to as the location of a distribution.

Memoryless Property of the Exponential Distribution

The exponential distribution has an interesting and unique property: it is memoryless. In other words,

$$P[X > x + s \mid X > x] = P[X > s].$$

To see this, note that

$$P[X > x] = \int_x^{\infty} f(t) dt = \int_x^{\infty} \lambda e^{-\lambda t} dt = e^{-\lambda x}.$$

Then

$$\begin{aligned} P[X > x + s \mid X > x] &= \frac{P[(X > x + s) \cap (X > x)]}{P[X > x]} = \frac{P[X > x + s]}{P[X > x]} \\ &= \frac{e^{-\lambda(x+s)}}{e^{-\lambda x}} = e^{-\lambda s} = P[X > s]. \end{aligned}$$

Time to Several Arrivals

The exponential distribution describes the time to the first (or next) arrival in a Poisson process.

Generalization: the time T_r needed for $r \in \mathbb{N} \setminus \{0\}$ arrivals to occur.

The cumulative distribution function is given by

$$\begin{aligned} F_{T_r}(t) &= P[T_r < t] \\ &= 1 - P[T_r > t] \\ &= 1 - P[\text{strictly less than } r \text{ arrivals before } t] \\ &= 1 - \sum_{n=0}^{r-1} \frac{(\lambda t)^n}{n!} e^{-\lambda t} \end{aligned}$$

for $t > 0$ and $F_{T_r}(t) = 0$ for $t < 0$.

Time to Several Arrivals

As before, we find, for $t \geq 0$,

$$\begin{aligned}f_{T_r}(t) &= F'_{T_r}(t) \\&= \lambda e^{-\lambda t} \sum_{n=0}^{r-1} \frac{(\lambda t)^n}{n!} - \lambda e^{-\lambda t} \sum_{n=1}^{r-1} \frac{(\lambda t)^{n-1}}{(n-1)!} \\&= \lambda e^{-\lambda t} \frac{(\lambda t)^{r-1}}{(r-1)!} \\&= \frac{\lambda^r}{(r-1)!} t^{r-1} e^{-\lambda t}\end{aligned}$$

and $f_{T_r}(t) = 0$ for $t < 0$.

The Gamma Distribution

6.5. Definition. Let $\alpha, \beta \in \mathbb{R}$, $\alpha, \beta > 0$. A continuous random variable $(X, f_{\alpha, \beta})$ with density

$$f_{\alpha, \beta}(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

is said to follow a **gamma distribution** with parameters α and β . Here

$$\Gamma(\alpha) = \int_0^\infty z^{\alpha-1} e^{-z} dz, \quad \alpha > 0,$$

is the **Euler gamma function**.

Hence, the time needed for the next r arrivals in a Poisson process with rate λ is determined by a Gamma distribution with parameters

$$\alpha = r$$

and

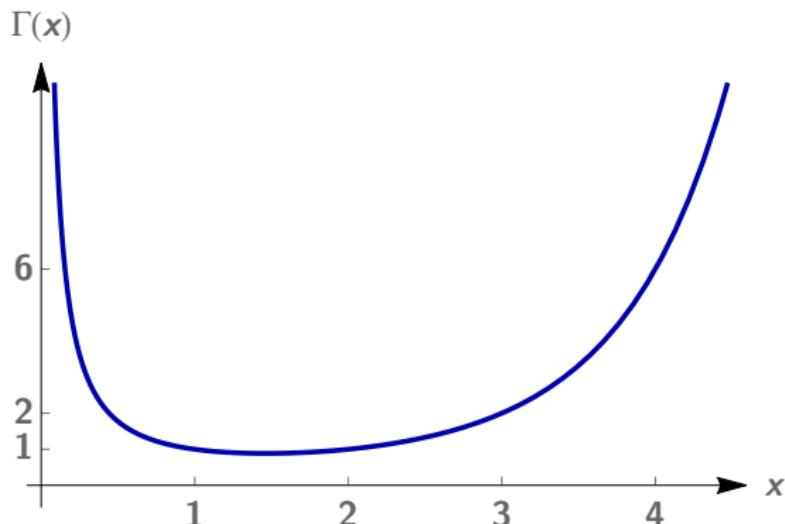
$$\beta = \lambda.$$

Gamma Distribution

The gamma function satisfies $\Gamma(1) = 1$ and $\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1)$ if $\alpha > 1$. In other words,

$$n! = \Gamma(n + 1) \quad \text{for } n \in \mathbb{N}.$$

Hence it is a continuous extension of the factorial function to the positive real numbers. Below is its graph for $\alpha \in (0, 5)$.



Mean, Variance, Moment-Generating Function

6.6. Theorem. Let $(X, f_{\alpha,\beta})$ be a Gamma distributed random variable with parameters $\alpha, \beta > 0$.

- (i) The moment-generating function of X is given by

$$m_X: (-\infty, \beta) \rightarrow \mathbb{R}, \quad m_X(t) = (1 - t/\beta)^{-\alpha}.$$

(ii) $E[X] = \alpha/\beta$.

(iii) $\text{Var}[X] = \alpha/\beta^2$.

Mean, Variance, Moment-Generating Function

Proof.

We will verify the moment-generating function only.

$$\begin{aligned}m_X(t) &= E[e^{tX}] = \int_0^\infty e^{tx} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x} dx \\&= \frac{\beta^\alpha}{\Gamma(\alpha)} \int_0^\infty x^{\alpha-1} e^{-x(\beta-t)} dx\end{aligned}$$

Substituting $y = x(\beta - t)$, we have $dy = (\beta - t)dx$ and

$$\begin{aligned}m_X(t) &= \frac{\beta^\alpha}{\Gamma(\alpha)} (\beta - t)^{-1} \int_0^\infty [y/(\beta - t)]^{\alpha-1} e^{-y} dy \\&= \frac{(\beta - t)^{-\alpha} \beta^\alpha}{\Gamma(\alpha)} \int_0^\infty y^{\alpha-1} e^{-y} dy \\&= (1 - t/\beta)^{-\alpha}.\end{aligned}$$



The Chi-Squared Distribution

The Gamma distribution is popular for modeling applications, since its parameters allow it to be fitted to many situations.

An important example is the **chi-squared distribution**.

6.7. Definition. Let $\gamma \in \mathbb{N}$. A continuous random variable (χ_{γ}^2, f_X) with density

$$f_{\gamma}(x) = \begin{cases} \frac{1}{\Gamma(\gamma/2)2^{\alpha}} x^{\gamma/2-1} e^{-x/2}, & x > 0, \\ 0, & x \leq 0, \end{cases}$$

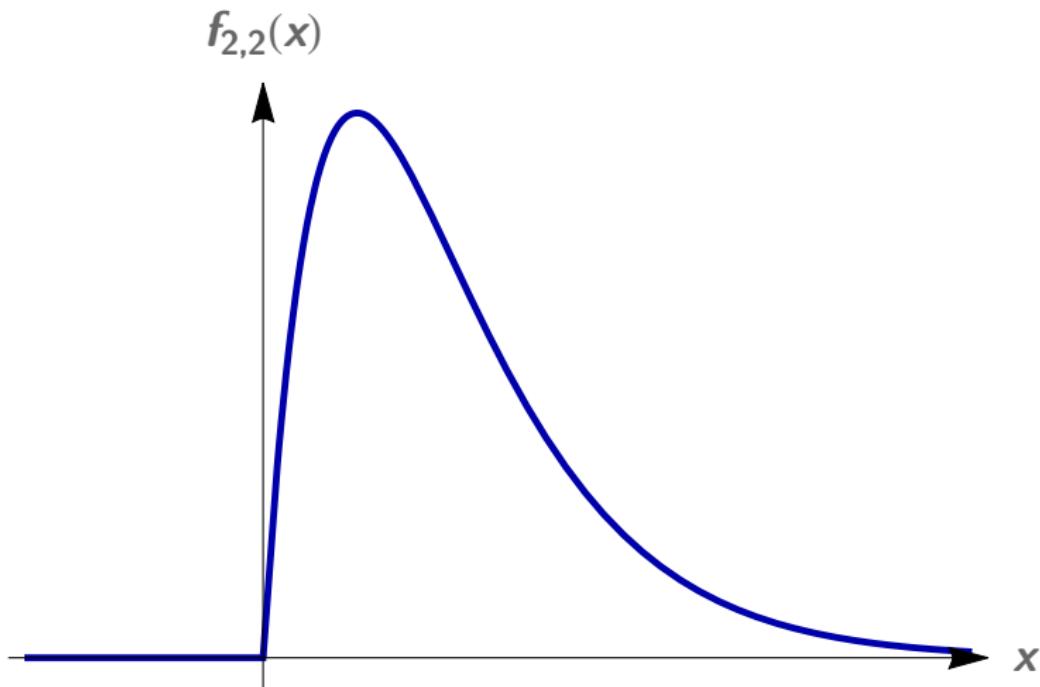
is said to follow a chi-squared distribution with γ **degrees of freedom**.

The chi-squared distribution is simply a gamma distribution with $\beta = 2$ and $\alpha = \gamma/2$. It is worth noting that

$$\mathbb{E}[\chi_{\gamma}^2] = \gamma, \quad \text{Var}[\chi_{\gamma}^2] = 2\gamma.$$

This distribution plays an important role in statistics.

Density of a Gamma Distribution with $\alpha = \beta = 2$



The Normal Distribution

Ceres

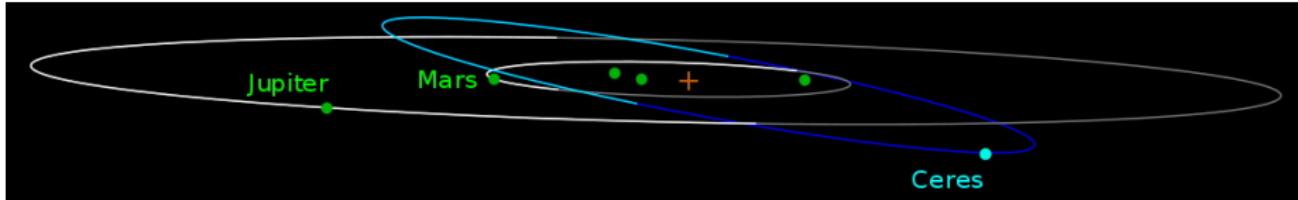


Portrait of Giuseppe Piazzi (1746-1826)
Bordiga, F. 1808. Smithsonian Institute Library.
File:Giuseppe Piazzi.jpg. (2013, November 2).
Wikimedia Commons, the free media repository.

On January 1st, 1801, the comet Ceres (later: planet, asteroid, dwarf planet) was discovered by the Italian priest Giuseppe Piazzi. He observed it 24 times until February 11th. When his observations were finally published in September 1801, Ceres could not be observed any more due to the sun's glare. So Piazzi's discovery could not be confirmed.

To find Ceres once it would become visible again at the end of the year, its position would need to be calculated.

But Piazzi had observed only around 1% of Ceres's orbit.



Orbit of Ceres File:Ceres Orbit.svg. (2016, January 12). Wikimedia Commons, the free media repository.

Carl Friedrich Gauß

The young mathematician Carl Friedrich Gauß heard about the Ceres problem and set to work. Within three months, he derived a prediction for the expected position of Ceres and published it in early December 1801. On December 31st, Ceres was found very close to the predicted position. This achievement of the 24-year-old Gauß established his reputation.

Gauß developed several new mathematical tools (such as the least-squares method).

But the most important idea was that a prediction would be impossible without understanding the mathematical function which described the errors and uncertainties in the Piazzi's observations. Starting from the premise that such a function exists in the first place, he derived the distribution which became known as the **Gaußian** or **normal distribution**.



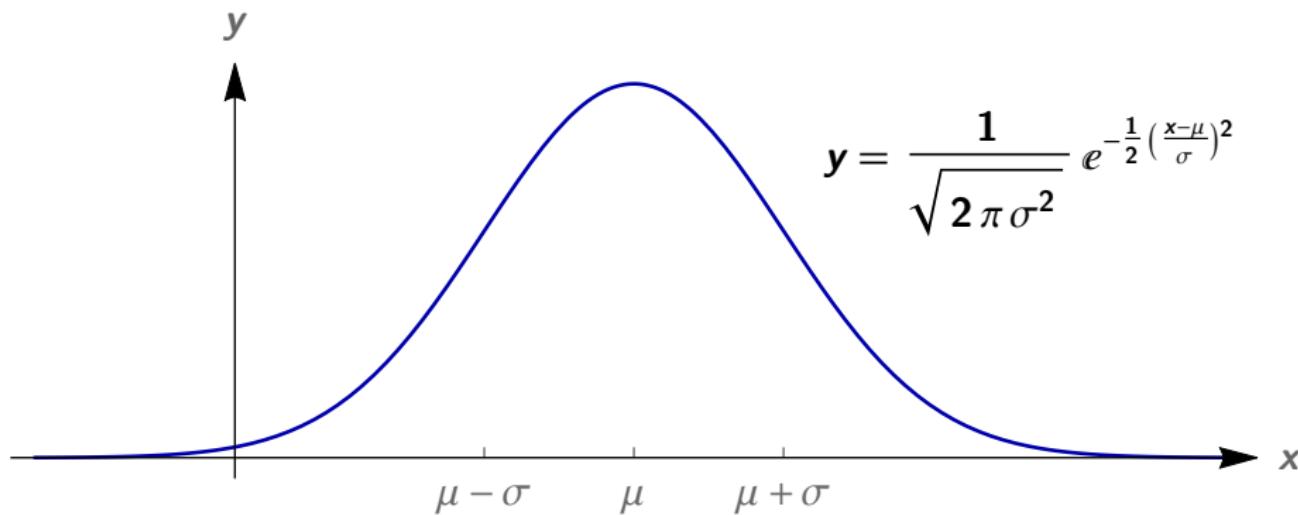
C. F. Gauß at 50 (1777-1855) Bendixson, S. D. 1828. Lithograph. Smithsonian Institute Library. File:Bendixen - Carl Friedrich Gauß, 1828.jpg. (2020, March 5). Wikimedia Commons, the free media repository.

Normal (Gauß) Distribution

7.1. Definition. Let $\mu \in \mathbb{R}$, $\sigma > 0$. A continuous random variable (X, f_X) with density

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

is said to follow a normal distribution with parameters μ and σ .



Normal Distribution

It is easily verified that $\int_{\mathbb{R}} f_X(x) dx = 1$ by using polar coordinates.

We write

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

whenever a random variable X follows a normal distribution with mean μ and variance σ^2 .

7.2. Theorem. Let (X, f_X) be a normally distributed random variable with parameters μ and σ .

(i) The moment-generating function of X is given by

$$m_X: \mathbb{R} \rightarrow \mathbb{R}, \quad m_X(t) = e^{\mu t + \sigma^2 t^2 / 2}.$$

(ii) $E[X] = \mu$.

(iii) $\text{Var}[X] = \sigma^2$.

Normal Distribution

Proof.

We will verify the moment-generating function only.

$$\begin{aligned}m_X(t) &= E[e^{tX}] = \int_{-\infty}^{\infty} \frac{e^{tx}}{\sqrt{2\pi}\sigma} e^{-((x-\mu)/\sigma)^2/2} dx \\&= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{tx - ((x-\mu)/\sigma)^2/2} dx\end{aligned}$$

We complete the square in the exponent to gain

$$tx - \frac{(x - \mu)^2}{2\sigma^2} = -\frac{(x - (\mu + \sigma^2 t))^2}{2\sigma^2} + \mu t + \sigma^2 t^2 / 2$$

Normal Distribution

Proof (continued).

Substituting into the integral,

$$\begin{aligned}m_X(t) &= \frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2} + \mu t + \sigma^2 t^2 / 2} dx \\&= e^{\mu t + \sigma^2 t^2 / 2} \underbrace{\frac{1}{\sqrt{2\pi}\sigma} \int_{-\infty}^{\infty} e^{-\frac{(x-(\mu+\sigma^2 t))^2}{2\sigma^2}} dx}_{=1}\end{aligned}$$

□

Standard Normal Distribution

7.3. Definition. A normally distributed random variable with parameters $\mu = 0$ and $\sigma = 1$ is called a **standard normal** random variable and denoted by Z .

The standard normal distribution is particularly important because any normally distributed random variable can be transformed into a standard-normally distributed one.

7.4. Theorem. Let X be a normally distributed random variable with mean μ and standard deviation σ . Then

$$Z := \frac{X - \mu}{\sigma}$$

has standard normal distribution.

Transformation of Random Variables

It is easily seen that $Z = \frac{X-\mu}{\sigma}$ has mean $E[Z] = 0$ and variance $\text{Var } Z = 1$, but it is not clear that Z is normally distributed. To see this, we need to find the density of Z .

Hence it is worth studying the density of transformed variables in general.

7.5. Theorem. Let X be a continuous random variable with density f_X . Let $Y = \varphi \circ X$, where $\varphi: \mathbb{R} \rightarrow \mathbb{R}$ is strictly monotonic and differentiable. The density for Y is then given by

$$f_Y(y) = f_X(\varphi^{-1}(y)) \cdot \left| \frac{d\varphi^{-1}(y)}{dy} \right| \quad \text{for } y \in \text{ran } \varphi$$

and

$$f_Y(y) = 0 \quad \text{for } y \notin \text{ran } \varphi.$$

Transformation of Random Variables

Proof.

We assume without loss of generality that φ is strictly decreasing. (The case where φ is strictly increasing is analogous.)

The cumulative distribution function for Y is given by

$$F_Y(y) = P[Y \leq y] = P[\varphi(X) \leq y].$$

Since φ is strictly decreasing, φ^{-1} exists and is also decreasing. Suppose that $y \in \text{ran } \varphi$. Then

$$\begin{aligned} F_Y(y) &= P[\varphi(X) \leq y] \\ &= P[\varphi^{-1}(\varphi(X)) \geq \varphi^{-1}(y)] \\ &= P[X \geq \varphi^{-1}(y)] \\ &= 1 - P[X \leq \varphi^{-1}(y)] \\ &= 1 - F_X(\varphi^{-1}(y)). \end{aligned}$$

Transformation of Random Variables

Proof (continued).

Since φ is strictly continuous, the range of φ is an interval in \mathbb{R} . If $y \notin \text{ran } \varphi$, then either $y > z$ for all $z \in \text{ran } \varphi$ or $y < z$ for all $z \in \text{ran } \varphi$.

We then see that

$$F_Y(y) = P[Y \leq y] = P[\varphi(X) \leq y] = \begin{cases} 0 & \text{if } y < z \text{ for all } z \in \text{ran } \varphi \\ 1 & \text{if } y > z \text{ for all } z \in \text{ran } \varphi \end{cases}$$

To obtain the density f_Y , we differentiate F_Y . For $y \in \text{ran } \varphi$ we have

$$\begin{aligned} f_Y(y) &= F'_Y(y) = -f_X(\varphi^{-1}(y)) \frac{d\varphi^{-1}(y)}{dy} \\ &= f_X(\varphi^{-1}(y)) \cdot \left| \frac{d\varphi^{-1}(y)}{dy} \right|. \end{aligned}$$

If $y \notin \text{ran } \varphi$, F_Y is constant and hence $f_Y = F'_Y = 0$. □

Standard Normal Distribution

We can now prove Theorem 7.4. We have $Z = \varphi \circ X$, where $\varphi(x) = \frac{x-\mu}{\sigma}$ is strictly increasing and differentiable with $\text{ran } \varphi = \mathbb{R}$. Note that

$$\varphi^{-1}(z) = \sigma z + \mu, \quad \frac{d\varphi^{-1}(z)}{dz} = \sigma > 0.$$

Using

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

we have

$$f_Z(z) = f_X(\varphi^{-1}(z)) \cdot \left| \frac{d\varphi^{-1}(z)}{dz} \right| = \frac{1}{\sqrt{2\pi}\sigma} e^{-(z-\mu)^2/2\sigma^2} \cdot \sigma = \frac{1}{\sqrt{2\pi}} e^{-z^2/2},$$

which is the density of the standard normal distribution. Hence the variable $Z = \frac{X-\mu}{\sigma}$ is standard normal.

Transforming Variables

We can verify Theorem 7.4 with Mathematica:

$$\text{TransformedDistribution}\left[\frac{x - \mu}{\sigma}, x \approx \text{NormalDistribution}[\mu, \sigma]\right]$$
$$\text{NormalDistribution}[0, 1]$$

Question. If X is standard normal, what is the density of X^2 ?

$$\text{PDF}[\text{TransformedDistribution}[x^2, x \approx \text{NormalDistribution}[0, 1]], x]$$
$$\begin{cases} \frac{e^{-x/2}}{\sqrt{2\pi}\sqrt{x}} & x > 0 \\ 0 & \text{True} \end{cases}$$

Note that the function $f(x) = x^2$ is not monotonic, so Theorem 7.5 can not be applied and a formal calculation needs to be done by hand!

Standard Normal Distribution

The cumulative distribution function of the standard normal distribution is often denoted by Φ ,

$$\Phi(z) := \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-t^2/2} dt.$$

The values of Φ are given in Table V of Appendix A. In Mathematica, the cumulative distribution function is expressed through the error function, defined as

$$\text{erf}(z) := \frac{2}{\sqrt{\pi}} \int_0^z e^{-t^2} dt, \quad \text{erfc}(z) := 1 - \text{erf}(z).$$

Hence,

```
CDF[NormalDistribution[0, 1], x]
```

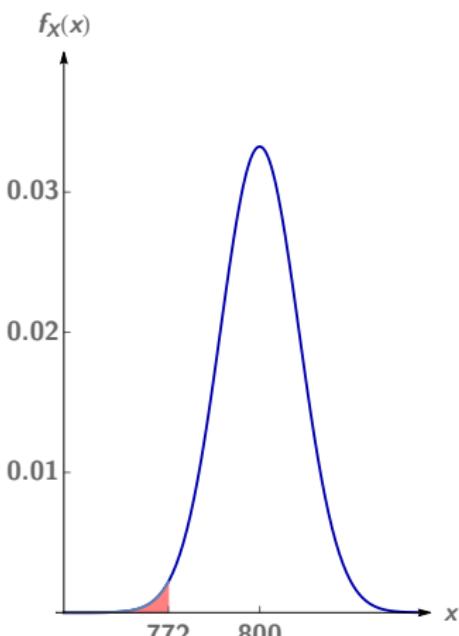
$$\frac{1}{2} \text{Erfc}\left[-\frac{x}{\sqrt{2}}\right]$$

Standard Normal Distribution

7.6. Example. The breaking strength of a synthetic fabric is denoted X , and it is normally distributed with mean $\mu = 800$ N and standard deviation $\sigma = 12$ N.

A purchaser of the fabric requires the fabric to have a strength of at least 772 N. A fabric sample is randomly selected and tested. To find $P[X \geq 772]$, we calculate

$$\begin{aligned} P[X < 772] &= P\left[\frac{X - \mu}{\sigma} < \frac{772 - 800}{12}\right] \\ &= P[Z < -2.33] \\ &= \Phi(-2.33) = 0.01. \end{aligned}$$



Hence the sample is 99% likely to pass inspection.

Standard Normal Distribution

7.7. Example. Let X denote the amount of radiation that can be absorbed by an individual before death ensues. Assume that X is normal with a mean dosage of 500 roentgens and a standard deviation of 150 roentgens. Above what dosage level will only 5% of those exposed survive?

Here we want to find x_0 such that $P[X \geq x_0] = 0.05$. Standardizing,

$$P[X \geq x_0] = P\left[\frac{X - 500}{150} \geq \frac{x_0 - 500}{150}\right] = P\left[Z \geq \frac{x_0 - 500}{150}\right] = 0.05$$

From Table V, $P[Z \geq 1.64] = 0.0505$ and $P[Z \geq 1.65] = 0.0495$.

Interpolating, we take $P[Z \geq 1.645] \approx 0.0500$, so we have

$$\frac{x_0 - 500 \text{ roentgen}}{150 \text{ roentgen}} = 1.645 \quad \Leftrightarrow \quad x_0 = 746.75 \text{ roentgen}.$$

Estimates on Variability

In general, the following estimates are often useful:

7.8. Theorem. Let X be normally distributed with parameters μ and σ .
Then

$$P[-\sigma < X - \mu < \sigma] = 0.68$$

$$P[-2\sigma < X - \mu < 2\sigma] = 0.95$$

$$P[-3\sigma < X - \mu < 3\sigma] = 0.997$$

Hence 68% of the values of a normal random variable lie within one standard deviation of the mean, 95% lie within two standard deviations, and 99.7% lie within three standard deviations. This rule of thumb will be especially important in statistics, where the number of “extraordinary” events needs to be judged.

Estimates on Variability

7.9. Example. Table of mean weights and heights of 12-month-old babies from a Chinese infant care book.

男童的标准

项 目	-2SD	中位数	+2SD
体重(kg)	8.1	10.2	12.4
身高(cm)	70.7	76.1	81.5

女童的标准

项 目	-2SD	中位数	+2SD
体重(kg)	7.4	9.5	11.6
身高(cm)	68.6	74.3	80.0

The Chebyshev Inequality

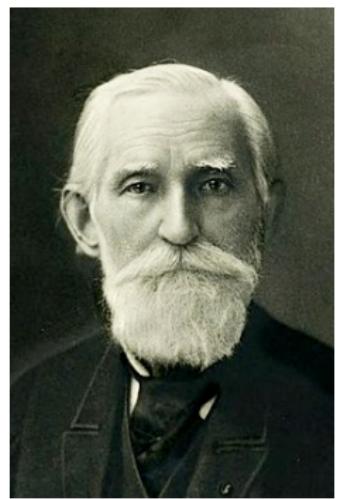
Let $c > 0$ be any real number. Then

$$\begin{aligned} E[X^2] &= \int_{-\infty}^{\infty} x^2 f_X(x) dx \geq \int_{|x| \geq c} x^2 f_X(x) dx \\ &\geq c^2 \int_{|x| \geq c} f_X(x) dx \\ &= c^2 \cdot P[|X| \geq c] \end{aligned}$$

More generally, for $k \in \mathbb{N} \setminus \{0\}$,

$$P[|X| \geq c] \leq \frac{E[|X|^k]}{c^k}. \quad (7.1)$$

This is one version of **Chebyshev's inequality**. The inequality also holds for discrete random variables, with an analogous proof.



Pafnuty Lvovich Chebyshev (1821–1894)
File:Pafnuty Lvovich Chebyshev.jpg. (2017, December 2). Wikimedia Commons, the free media repository.

Variability Estimate from Chebyshev's Inequality

If we replace X with $X - \mu$ in (7.1) and set $k = 2$ and $c = m \cdot \sigma$, $m > 0$, we obtain the estimate

$$P[|X - \mu| \geq m\sigma] \leq \frac{1}{m^2}. \quad (7.2)$$

or, equivalently,

$$P[-m\sigma < X - \mu < m\sigma] \geq 1 - \frac{1}{m^2} \quad (7.3)$$

Comparing (7.2) with Theorem 7.8, we see that the estimates in the theorem are tighter.

This is not surprising, as Chebyshev's rule is valid for any random variable with finite second moment, while the previous theorem uses the specific properties of the normal distribution.

(Im-)Practical Application of Chebyshev's Inequality

7.10. Example. From an analysis of company records, a materials control manager estimates that the mean and standard deviation of the "lead time" required in ordering a small valve are 8 days and 1.5 days, respectively. She does not know the distribution of the lead time, but she is willing to assume the estimates of the mean and standard deviation to be absolutely correct.

The manager would like to determine a time interval such that the probability is at least $8/9$ that the order will be received during that time. That is,

$$1 - \frac{1}{k^2} = \frac{8}{9},$$

so that $k = 3$ and $\mu \pm k\sigma = (8 \pm 4.5)$ days.

This interval may well be too large to be of any value to the manager, in which case she may elect to learn more about the distribution of lead times.

Approximating the Binomial Distribution



Abraham De Moivre (1667-1754).
Portrait. Faber. 1736.
File:Abraham de moivre.jpg.
(2015, October 30). Wikimedia
Commons, the free media
repository.



**Pierre-Simon de Laplace
(1749-1827).** Engraving.
File:Pierre-Simon-Laplace
(1749-1827).jpg. (2017, June 6).
Wikimedia Commons, the free
media repository.

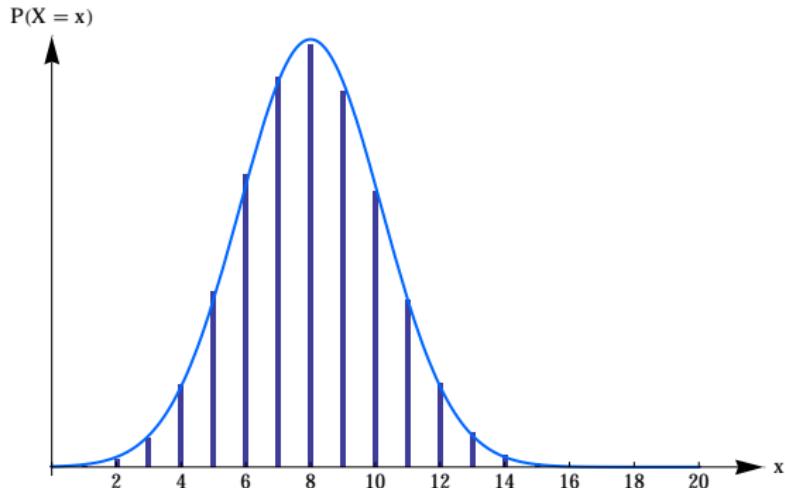
Long before Gauß discovered the normal distribution in 1801, it had been published 60 years earlier, in 1738. De Moivre had wanted to approximate the shape of the binomial distribution, considering the behavior of 3600 coin tosses. In 1810, Laplace proved the general result for $0 < p < 1$.

7.11. Theorem of De Moivre-Laplace. Denote by S_n the number of successes in a sequence of n i.i.d. Bernoulli trials with probability of success $0 < p < 1$. Then

$$\lim_{n \rightarrow \infty} P\left[a < \frac{S_n - np}{\sqrt{np(1-p)}} \leq b\right] = \frac{1}{\sqrt{2\pi}} \int_a^b e^{-x^2/2} dx.$$

Approximating the Binomial Distribution

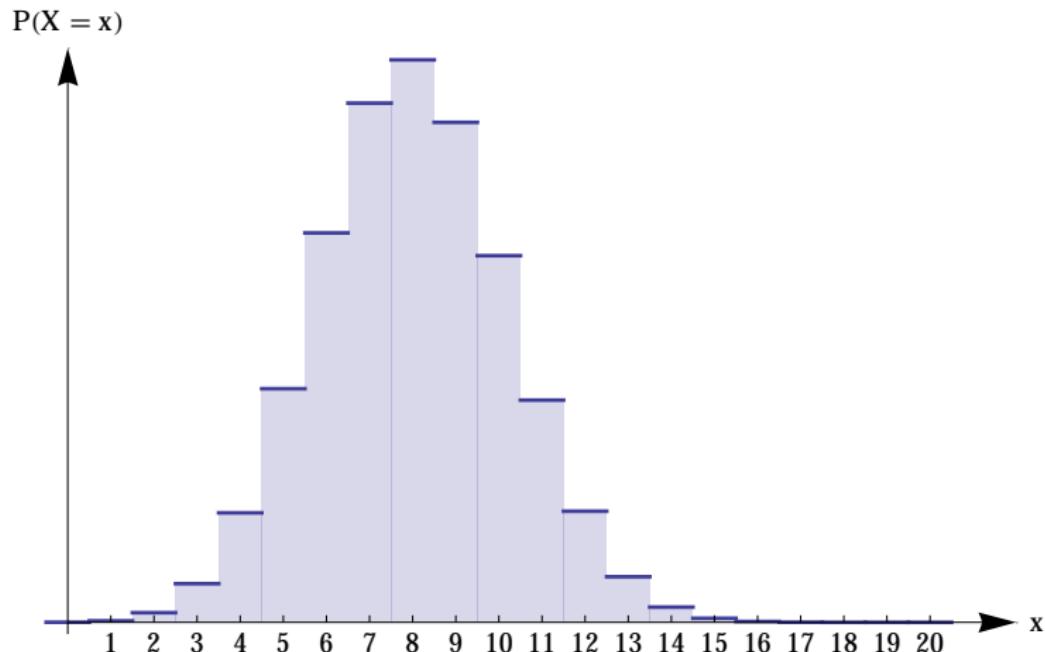
Intuitively, for large n , the binomial distribution with parameters n and p behaves as a normal distribution with mean $\mu = np$ and variance $\sigma^2 = npq$. This is illustrated below for $n = 20$ and $p = 0.4$:



The height of the vertical bars represents the values of $P[X = x]$ according to the binomial distribution, while the density curve of the corresponding normal distribution has been superimposed.

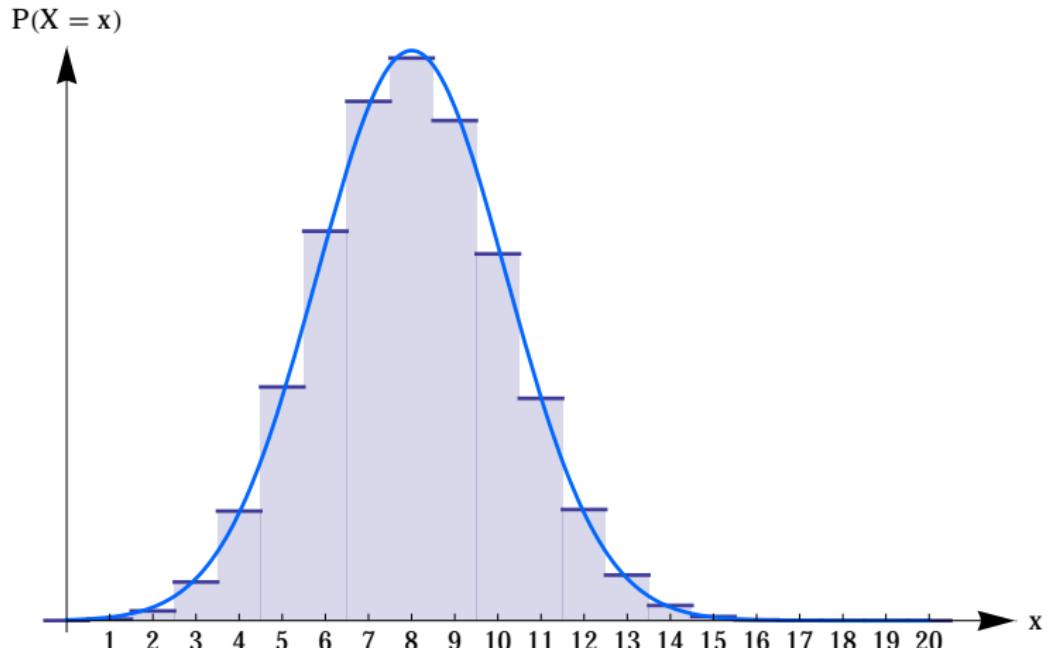
Approximating the Binomial Distribution

We would like to use the normal distribution to approximate the cumulative distribution function of the binomial distribution.



Approximating the Binomial Distribution

It is clear that for each $y = 0, \dots, 20$ the sum over all $x \leq y$ corresponds to the area of the bars to the left of y . Superimposing the normal distribution, we see that we can approximate this sum by integrating to $y + 1/2$:



Approximation and the Half-Unit Correction

Hence, for $y = 0, \dots, n$,

$$P[X \leq y] = \sum_{x=0}^y \binom{n}{x} p^x (1-p)^{n-x} \approx \Phi\left(\frac{y + 1/2 - np}{\sqrt{np(1-p)}}\right).$$

This additional term $1/2$ is known as the **half-unit correction** for the normal approximation to the cumulative binomial distribution function. It is necessary because in practice we do not have the limit $n \rightarrow \infty$ but rather a finite value of n , which may not even be especially large.

This approximation is good if p is close to $1/2$ and $n > 10$. Otherwise, we require that

$$np > 5 \quad \text{if } p \leq 1/2 \quad \text{or} \quad n(1-p) > 5 \quad \text{if } p > 1/2.$$

Approximating the Binomial Distribution

7.12. Example. In sampling from a production process that produces items of which 20% are defective, a random sample of 100 items is selected each hour of each production shift. The number of defectives in a sample is denoted by X .

To find, say, $P[X \leq 15]$ we might use the normal approximation as follows:

$$\begin{aligned} P[X \leq 15] &\approx P\left[Z \leq \frac{15 - 100 \cdot 0.2}{\sqrt{100 \cdot 0.2 \cdot 0.8}}\right] = P[Z \leq -1.25] \\ &= \Phi(-1.25) = 0.1056 \end{aligned}$$

The half-unit correction would instead give

$$P[X \leq 15] \approx P\left[Z \leq \frac{15.5 - 20}{4}\right] = 0.130$$

The correct result is $P[X \leq 15] = \sum_{k=0}^{15} \binom{100}{k} 0.2^k 0.8^{100-k} = 0.1285$.

Lyapunov's Central Limit Theorem



Aleksandr Mikhailovich Lyapunov
(1857-1918). File:Alexander Ljapunow
Jung.jpg. (2017, January 29).
Wikimedia Commons, the free media
repository.

Today there exist various “Central Limit Theorems” that generalize the Theorem of De Moivre - Laplace. The following is due to Lyapunov and was established a century after Laplace’s work.

7.13. Central Limit Theorem. Let (X_i) be a sequence of independent, but not necessarily identical random variables whose third moments exist and satisfy a certain technical condition.

Let

$$Y_n = X_1 + \cdots + X_n.$$

Then for any $z \in \mathbb{R}$,

$$P\left[\frac{Y_n - E[Y_n]}{\sqrt{\text{Var}[Y_n]}} \leq z\right] \xrightarrow{n \rightarrow \infty} \frac{1}{\sqrt{2\pi}} \int_{-\infty}^z e^{-x^2/2} dx.$$

Experimental Error

Lyapunov's Central Limit theorem is at the core of the belief by experimentalists that "random error" may be described by the normal distribution. The idea is that any "random disturbance" of a measurement is the sum of many inscrutable and random effects which individually cannot be tracked. However, their sum will be well-described by the normal distribution.

The French physicist Gabriel Lippman wrote to Henri Poincaré:

Tout le monde y croit cependant, car les expérimentateurs s'imaginent que c'est un théorème de mathématiques, et les mathématiciens que c'est un fait expérimental.

"Everybody believes in the exponential law of errors: the experimenters, because they think it can be proved by mathematics; and the mathematicians, because they believe it has been established by observation."

Multivariate Random Variables

Multivariate Random Variables

Often, a single random variable is not sufficient to describe a physical problem. This may, for example, be the case when we are interested in the effect of one random quantity on another. In such a case we consider two (or more) random variables together.

Formally, we then define a “vector” of which each component is itself a (“scalar”) random variable.

We call such a vector a ***random vector*** or a ***multi-variate random variable*** or an ***n-dimensional random variable***. The components can be discrete or continuous random variables, and even mixtures of the two.

In this section we will for the most part focus on bivariate (two-dimensional) random variables where either both components are discrete or both components are continuous random variables.

Discrete Multivariate Random Variables

8.1. Definition. Let S be a sample space and Ω a countable subset of \mathbb{R}^n . A **discrete multivariate random variable** is a map

$$\mathbf{X}: S \rightarrow \Omega$$

together with a function $f_{\mathbf{X}}: \Omega \rightarrow \mathbb{R}$ with the properties that

- (i) $f_{\mathbf{X}}(x) \geq 0$ for all $x = (x_1, \dots, x_n) \in \Omega$ and
- (ii) $\sum_{x \in \Omega} f_{\mathbf{X}}(x) = 1$.

The function $f_{\mathbf{X}}$ is called the **joint density function** of the random variable \mathbf{X} .

Discrete Multivariate Random Variables

We consider the multivariate random variable \mathbf{X} to have n components, i.e.,

$$\mathbf{X} = (X_1, \dots, X_n).$$

We often write

$$f_{\mathbf{X}}(x_1, \dots, x_n) = f_{X_1 \dots X_n}(x_1, \dots, x_n)$$

The joint density function $f_{\mathbf{X}}$ gives the probability that the tuple (X_1, \dots, X_n) assumes a given value $x \in \mathbb{R}^n$, i.e.,

$$f_{\mathbf{X}}(x_1, \dots, x_n) = P[X_1 = x_1 \text{ and } X_2 = x_2 \text{ and } \dots \text{ and } X_n = x_n].$$

Given two random variables, we may write (X, Y) instead of (X_1, X_2) and use similar notation for three or larger numbers of components.

Discrete Bivariate Random Variables

8.2. Example. Suppose we roll two six-sided dice, obtaining results (i, j) with $i, j = 1, \dots, 6$. Let us define

$$X := i + j \bmod 5, \quad Y = i - j \bmod 5.$$

Then we can find the values of the probability density function by Cardano's rule. The number of outcomes leading to each event (X, Y) is

x/y	0	1	2	3	4
0	1	1	4	1	1
1	1	2	1	2	1
2	2	1	1	1	2
3	2	1	1	1	2
4	1	2	1	2	1

so each number in the table must be divided by 36 to obtain the corresponding probability. For example, $P[X = 1 \text{ and } Y = 1] = 1/18$.

Marginal Density

While each element of the table gives us $36 \cdot P[X = x \text{ and } Y = y]$, we can find the probability of the event $X = x$ by adding up all relevant probabilities:

$$P[X = x] = \sum_{y=0}^4 P[X = x \text{ and } Y = y]$$

x/y	0	1	2	3	4
0	1	1	4	1	1
1	1	2	1	2	1
2	2	1	1	1	2
3	2	1	1	1	2
4	1	2	1	2	1

For example,

$$P[X = 0] = (1 + 1 + 4 + 1 + 1) / 36 = 8/36.$$

This procedure can be justified by considering the corresponding event in the sample space.

By summing in this way, we can determine $P[X = x]$ for all x . This is called the **marginal density** for X .

Marginal Density of a Discrete Random Variable

8.3. Definition. Let $(\mathbf{X}, f_{\mathbf{X}})$ be a discrete multivariate random variable. We define the **marginal density** f_{x_k} for X_k , $k = 1, \dots, n$, by

$$f_{X_k}(x_k) = \sum_{x_1, \dots, x_{k-1}, x_{k+1}, \dots, x_n} f_{\mathbf{X}}(x_1, \dots, x_n).$$

8.4. Example.

x/y	0	1	2	3	4	$f_X(x)$
0	1	1	4	1	1	$8/36$
1	1	2	1	2	1	$7/36$
2	2	1	1	1	2	$7/36$
3	2	1	1	1	2	$7/36$
4	1	2	1	2	1	$7/36$
$f_Y(y)$	$7/36$	$7/36$	$8/36$	$7/36$	$7/36$	1

Independence of two Random Variables

Question. Considering the table:

x/y	0	1	2	3	4	$f_X(x)$
0	1	1	4	1	1	$8/36$
1	1	2	1	2	1	$7/36$
2	2	1	1	1	2	$7/36$
3	2	1	1	1	2	$7/36$
4	1	2	1	2	1	$7/36$
$f_Y(y)$	$7/36$	$7/36$	$8/36$	$7/36$	$7/36$	1

Do you think that X and Y are independent?

- ▶ Yes
- ▶ No
- ▶ It's not possible to tell from the table.

Independence of Random Variables

If $(\mathbf{X}, f_{\mathbf{X}})$ is a discrete **bivariate** random variable, i.e., $\mathbf{X} = (X_1, X_2)$, we say that X_1 and X_2 are **independent** if

$$P[X_1 = x_1 \text{ and } X_2 = x_2] = P[X_1 = x_1] \cdot P[X_2 = x_2].$$

In other words, if

$$f_{\mathbf{X}}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2).$$

(The joint density is the product of the marginal densities.)

It is possible to generalize this in the obvious (but notationally cumbersome) way to n -variate random variables.

We will mostly be interested in cases where $\mathbf{X} = (X_1, \dots, X_n)$ and all the components are independent, i.e.,

$$f_{\mathbf{X}}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

Independence of two Random Variables

8.5. Example.

x/y	0	1	2	3	4	$f_X(x)$
0	1	1	4	1	1	$8/36$
1	1	2	1	2	1	$7/36$
2	2	1	1	1	2	$7/36$
3	2	1	1	1	2	$7/36$
4	1	2	1	2	1	$7/36$
$f_Y(y)$	$7/36$	$7/36$	$8/36$	$7/36$	$7/36$	1

The variables X and Y are not independent since, for example,

$$P[X = 1 \text{ and } Y = 1] = 1/18$$

but

$$P[X = 1] \cdot P[Y = 1] = \frac{7}{36} \cdot \frac{7}{36}$$

and the two expressions are not equal.

Conditional Density

Suppose that $(\mathbf{X}, f_{\mathbf{X}})$ is a discrete bivariate random variable, i.e., $\mathbf{X} = (X_1, X_2)$, and that X_2 is known to have taken on a certain value.

Then, applying elementary probability laws,

$$P[X_1 = x_1 \mid X_2 = x_2] = \frac{P[X_1 = x_1 \text{ and } X_2 = x_2]}{P[X_2 = x_2]} = \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_2}(x_2)}.$$

We hence define the ***conditional density***

$$f_{X_1 \mid X_2}(x_1) := \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad \text{whenever } f_{X_2}(x_2) > 0,$$

where f_{X_2} is the marginal density of X_2 .

Continuous Random Variables

8.6. Definition. Let S be a sample space. A ***continuous multivariate random variable*** is a map

$$\mathbf{X}: S \rightarrow \mathbb{R}^n$$

together with a function $f_{\mathbf{X}}: \mathbb{R}^n \rightarrow \mathbb{R}$ with the properties that

- (i) $f_{\mathbf{X}}(x) \geq 0$ for all $x = (x_1, \dots, x_n) \in \mathbb{R}^n$ and
- (ii) $\int_{\mathbb{R}^n} f_{\mathbf{X}}(x) dx = 1.$

The function $f_{\mathbf{X}}$ is called the ***joint density function*** of the random variable \mathbf{X} .

Continuous Random Variables

The integral of $f_{\mathbf{X}}$ is interpreted as the probability that \mathbf{X} assumes values in a given domain $\Omega \subset \mathbb{R}^n$,

$$P[\mathbf{X} \in \Omega] = \int_{\Omega} f_{\mathbf{X}}(x) dx.$$

For example, if $\mathbf{X} = (X_1, X_2)$,

$$P[a \leq X_1 \leq b \text{ and } c \leq X_2 \leq d] = \int_a^b \int_c^d f_{X_1 X_2}(x_1, x_2) dx_1 dx_2$$

for $a \leq b$, $c \leq d$.

But of course non-rectangular domains can be considered as well.

We now make definitions for continuous random variables that are completely analogous to those for the discrete case.

Continuous Multivariate Random Variables

We define the **marginal density** of X_k , $k = 1, \dots, n$, by

$$f_{X_k}(x_k) = \int_{\mathbb{R}^{n-1}} f_{\mathbf{X}}(x) dx_1 \dots dx_{k-1} dx_{k+1} \dots dx_n.$$

We say that two continuous random variables are **independent** if

$$f_{\mathbf{X}}(x_1, x_2) = f_{X_1}(x_1) \cdot f_{X_2}(x_2).$$

and we are often interested in the case where a full set of n components of a multivariate random variable is independent:

$$f_{\mathbf{X}}(x_1, \dots, x_n) = f_{X_1}(x_1) \cdots f_{X_n}(x_n).$$

The **conditional density** for continuous bivariate random variables is similarly

$$f_{X_1|X_2}(x_1) := \frac{f_{X_1 X_2}(x_1, x_2)}{f_{X_2}(x_2)} \quad \text{whenever } f_{X_2}(x_2) > 0.$$

Expectation

We define the **expected value** or **expectation** for \mathbf{X} as the vector

$$\mathbb{E}[\mathbf{X}] = \begin{pmatrix} \mathbb{E}[X_1] \\ \vdots \\ \mathbb{E}[X_n] \end{pmatrix}$$

where $\mathbb{E}[X_k]$ is calculated using the marginal density of X_k , $k = 1, \dots, n$,

$$\mathbb{E}[X_k] = \sum_{x_k} x_k f_{X_k}(x_k) = \sum_{x \in \Omega} x_k f_{\mathbf{X}}(x)$$

and

$$\mathbb{E}[X_k] = \int_{\mathbb{R}} x_k f_{X_k}(x_k) dx_k = \int_{\mathbb{R}^n} x_k f_{\mathbf{X}}(x) dx$$

for discrete and continuous random variables, respectively.

Expectation for Discrete Bivariate Random Variables

8.7. Example.

x/y	0	1	2	3	4	$f_X(x)$
0	1	1	4	1	1	$8/36$
1	1	2	1	2	1	$7/36$
2	2	1	1	1	2	$7/36$
3	2	1	1	1	2	$7/36$
4	1	2	1	2	1	$7/36$
$f_Y(y)$	$7/36$	$7/36$	$8/36$	$7/36$	$7/36$	1

$$E[X] = \sum_{(x,y) \in \Omega} x \cdot f_{XY}(x, y) = \sum_{x=0}^4 x \cdot f_X(x) = \frac{70}{36}$$

$$E[Y] = \sum_{(x,y) \in \Omega} y \cdot f_{XY}(x, y) = \sum_{y=0}^4 y \cdot f_Y(y) = 2$$

Expectation for Functions of Random Vectors

Suppose $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}$ is a continuous function. Then

$$\varphi \circ \mathbf{X}: S \rightarrow \mathbb{R}$$

defines a scalar random variable. It is possible to prove that in this case,

$$E[\varphi \circ \mathbf{X}] = \sum_{x \in \Omega} \varphi(x) f_{\mathbf{X}}(x), \quad \text{or} \quad E[\varphi \circ \mathbf{X}] = \int_{\mathbb{R}^n} \varphi(x) f_{\mathbf{X}}(x) dx.$$

For $\varphi(x_1, \dots, x_n) = x_k$ we regain the definition of $E[X_k]$.

Expectation for the Sum of Two Random Variables

8.8. Remark. If (X, Y) is a discrete bivariate random variable and $\varphi(x, y) = x + y$, we have

$$\begin{aligned} E[X + Y] &= \sum_{(x,y) \in \Omega} (x + y) \cdot f_{XY}(x, y) \\ &= \sum_{(x,y) \in \Omega} x \cdot f_{XY}(x, y) + \sum_{(x,y) \in \Omega} y \cdot f_{XY}(x, y) \\ &= E[X] + E[Y]. \end{aligned}$$

This establishes the addition property of the expectation that we introduced earlier.

An analogous calculation may be used for continuous random variables.

Variance and Covariance for Bivariate Random Variables

Let us calculate the variance of the sum of two random variables:

$$\begin{aligned}\text{Var}[X + Y] &= E[((X + Y) - E[X + Y])^2] \\&= E[((X - E[X]) + (Y - E[Y]))^2] \\&= E[(X - E[X])^2 + (Y - E[Y])^2 + 2(X - E[X])(Y - E[Y])] \\&= \text{Var}[X] + \text{Var}[Y] + 2E[(X - E[X])(Y - E[Y])]\end{aligned}\quad (8.1)$$

In general,

$$\text{Var}[X + Y] \neq \text{Var}[X] + \text{Var}[Y].$$

We define the **covariance of (X, Y)** ,

$$\text{Cov}[X, Y] = E[(X - \mu_X)(Y - \mu_Y)],$$

where we have used μ to denote the expectations. Note that

$$\text{Cov}[X, Y] = \text{Cov}[Y, X] \quad \text{and} \quad \text{Cov}[X, X] = \text{Var}[X].$$

The Covariance Matrix

For a multivariate random variable \mathbf{X} we define the **covariance matrix**

$$\text{Var}[\mathbf{X}] = \begin{pmatrix} \text{Var}[X_1] & \text{Cov}[X_1, X_2] & \dots & \text{Cov}[X_1, X_n] \\ \text{Cov}[X_1, X_2] & \text{Var}[X_2] & \ddots & \vdots \\ \vdots & \ddots & \ddots & \text{Cov}[X_{n-1}, X_n] \\ \text{Cov}[X_1, X_n] & \dots & \text{Cov}[X_{n-1}, X_n] & \text{Var}[X_n] \end{pmatrix}.$$

It is possible to prove (through tedious calculation) that

$$\text{Var}[C\mathbf{X}] = C \text{Var}[\mathbf{X}] C^T$$

where $C \in \text{Mat}(n \times n; \mathbb{R})$ is a constant $n \times n$ matrix with real coefficients.

Covariance and Independence

Just as for the variance, a direct calculation yields

$$\text{Cov}[X, Y] = E[XY] - E[X]E[Y].$$

Furthermore, if two continuous random variables X and Y are independent, then $f_{XY}(x, y) = f_X(x)f_Y(y)$ and

$$\begin{aligned} E[XY] &= \iint_{\mathbb{R}^2} xy \cdot f_{XY}(x, y) dx dy \\ &= \iint_{\mathbb{R}^2} xy \cdot f_X(x)f_Y(y) dx dy \\ &= \left(\int_{\mathbb{R}} x \cdot f_X(x) dx \right) \left(\int_{\mathbb{R}} y \cdot f_Y(y) dy \right) \\ &= E[X]E[Y] \end{aligned}$$

Covariance and Independence

An analogous calculation works for discrete random variables. We have hence proved:

- ▶ If X and Y are independent, then $\text{Cov}[X, Y] = 0$.

However, the converse is not true:

- ▶ If $\text{Cov}[X, Y] = 0$, then X and Y are ***not necessarily independent.***

We note that we have also established that

$$\text{Var}[X + Y] = \text{Var}[X] + \text{Var}[Y]$$

if the random variables are independent.

The covariance is hence related to the independence of X and Y . However, it is not a measure for dependence, since two dependent variables can still have a vanishing covariance.

So we ask: what does the covariance actually measure?

Standardizing Random Variables

We note that the covariance scales with X and Y , i.e., if X and Y take on numerically large values, then the covariance will be large, while if X and Y take on small values, the covariance will be small. Therefore, by itself it does not serve very well as a measure of any fundamental properties of X and Y .

The solution is to standardize the random variables,

$$\tilde{X} := \frac{X - \mu_X}{\sigma_X}$$

is the standardized variable for X (assuming that both mean and variance of X exist and $\sigma_X \neq 0$).

Recall that

$$E[\tilde{X}] = 0,$$

$$\text{Var}[\tilde{X}] = 1.$$

The Pearson Correlation Coefficient



Karl Pearson (1857-1936) in 1912. File:Karl Pearson, 1912.jpg. (2016, January 17).
Wikimedia Commons, the free media repository.

Instead of $\text{Cov}[X, Y]$ we now consider

$$\begin{aligned}\text{Cov}[\tilde{X}, \tilde{Y}] &= E[\tilde{X} \tilde{Y}] - E[\tilde{X}] E[\tilde{Y}] \\ &= \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}\end{aligned}$$

The right-hand side is now scale-independent and unit-less (if X and Y have units).

This quotient is known as the **Pearson coefficient of correlation** of (X, Y) and denoted

$$\rho_{XY} := \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}.$$

Properties of the Correlation Coefficient

It can be shown that ρ_{XY} has the following properties

- (i) $-1 \leq \rho_{XY} \leq 1$,
- (ii) $|\rho_{XY}| = 1$ if and only if there exist numbers $\beta_0, \beta_1 \in \mathbb{R}$, $\beta_1 \neq 0$, such that

$$Y = \beta_0 + \beta_1 X$$

almost surely.

The proof is best performed in a vector-space setting, which we omit here.

The above properties give us a clue as to how the correlation coefficient might be interpreted: if it has modulus one, then X and Y are in a deterministically linear relationship. Let us therefore start from that angle.

Measuring Linearity of X and Y

Suppose that X and Y are related in a linear fashion, say

$$Y = \beta_0 + \beta_1 X, \quad (8.2)$$

with $\beta_1 \neq 0$. Then

$$\mu_Y = \beta_0 + \beta_1 \mu_X$$

and $\text{Var}[Y] = \beta_1^2 \text{Var}[X]$, so

$$\sigma_Y = |\beta_1| \sigma_X.$$

Measuring Linearity of X and Y

Using the standardized variables, we find that

$$\begin{aligned}\tilde{Y} &= \frac{Y - \mu_Y}{\sigma_Y} \\ &= \frac{\beta_0 + \beta_1 X - (\beta_0 + \beta_1 \mu_X)}{|\beta_1| \sigma_X} \\ &= \frac{\beta_1}{|\beta_1|} \frac{X - \mu_X}{\sigma_X} \\ &= \frac{\beta_1}{|\beta_1|} \tilde{X}.\end{aligned}$$

We conclude that X and Y are in a linear relationship if and only if the standardized variables are either equal or the negative of each other.

Measuring Linearity of X and Y

We now know that X and Y are deterministically linearly related if and only if

$$\tilde{X} + \tilde{Y} = 0 \quad \text{or} \quad \tilde{X} - \tilde{Y} = 0.$$

In order to measure in how far X and Y are not linearly related, it makes sense to consider the standard deviation of $\tilde{X} + \tilde{Y}$ and $\tilde{X} - \tilde{Y}$. If either of these were zero, the relationship would be deterministically linear.

We calculate

$$\text{Var}[\tilde{X} + \tilde{Y}] = \text{Var}[\tilde{X}] + \text{Var}[\tilde{Y}] + 2 \text{Cov}[\tilde{X}, \tilde{Y}] = 2 + 2\rho_{XY},$$

$$\text{Var}[\tilde{X} - \tilde{Y}] = \text{Var}[\tilde{X}] + \text{Var}[\tilde{Y}] - 2 \text{Cov}[\tilde{X}, \tilde{Y}] = 2 - 2\rho_{XY}.$$

If either of these two variances is small, then \tilde{X} and \tilde{Y} are “nearly proportional” and so X and Y are “nearly linearly” related.

The Fisher Transformation

In order to capture both of these positive quantities in a single manner, let us consider their quotient,

$$\sqrt{\frac{\text{Var}[\tilde{X} + \tilde{Y}]}{\text{Var}[\tilde{X} - \tilde{Y}]}} = \sqrt{\frac{1 + \rho_{XY}}{1 - \rho_{XY}}} \in (0, \infty)$$

If X and Y are linearly related, then this quotient will be either very small or very large.

It is “mathematically nicer” to take the logarithm:



Ronald Fisher (1890-1962) In 1913.
 File:Youngronaldfisher2.JPG. (2018, July 7).
 Wikimedia Commons, the free media repository.

$$\ln\left(\sqrt{\frac{\text{Var}[\tilde{X} + \tilde{Y}]}{\text{Var}[\tilde{X} - \tilde{Y}]}}\right) = \frac{1}{2} \ln\left(\frac{1 + \rho_{XY}}{1 - \rho_{XY}}\right) = \text{Artanh}(\rho_{XY}) \in \mathbb{R}. \quad (8.3)$$

This is known as the **Fisher transformation** of ρ_{XY} .

Positive and Negative Correlation

It follows that

$$\rho_{XY} = \tanh\left(\ln\left(\frac{\sigma_{\tilde{X}+\tilde{Y}}}{\sigma_{\tilde{X}-\tilde{Y}}}\right)\right).$$

- If $\rho_{XY} > 0$, then $\text{Var}[\tilde{X} + \tilde{Y}] > \text{Var}[\tilde{X} - \tilde{Y}]$, which implies that the relationship between X and Y is closer to $\tilde{X} = \tilde{Y}$ than to $\tilde{X} = -\tilde{Y}$. Hence, if X is large, Y tends to be large also.

We say that X and Y are **positively correlated**.

- If $\rho_{XY} < 0$, then $\text{Var}[\tilde{X} + \tilde{Y}] < \text{Var}[\tilde{X} - \tilde{Y}]$ and the situation is reversed. If X is large, Y tends to be small.

We say that X and Y are **negatively correlated**.

Since X and Y are still *random* variables, a large value of X only indicates a tendency for Y to be large/small but doesn't guarantee this. The closer ρ_{XY} is to ± 1 , the more pronounced these effects are.

The Bivariate Normal Distribution

8.9. Example. Suppose two random variables X and Y should each follow a (marginal) normal distribution, but are not independent.

The most common model is the so-called **bivariate normal distribution**, with density function

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} e^{-\frac{1}{2(1-\varrho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\varrho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right]}$$

where $-1 < \varrho < 1$.

The marginal distributions can be shown to be normal, $\mu_X = E[X]$, $\sigma_X^2 = \text{Var } X$ (and similarly for Y) and $\varrho = \rho_{XY}$ is indeed the correlation coefficient of X and Y .

Furthermore, X and Y are independent if and only if $\varrho = 0$.

This distribution will be discussed in detail in the assignments.

The Hypergeometric Distribution

Drawing Balls from an Urn

The classical example of a sequence of non-independent trials involves drawing colored balls from a box, traditionally called an *urn*.

Suppose that an urn contains a total of N balls, of which r are red balls and $N - r$ are black balls. We draw a sample of n balls from the urn. We **do not replace** each ball after drawing it.

The random variable X describes the number of red balls in our sample.

If we were to replace each ball after drawing, X would follow a binomial distribution. But now the probability of drawing a red ball depends on the previous outcomes.

Drawing Balls from an Urn

Given the number of objects N , the sample size n and the number r of red balls, we can apply Cardano's principle to calculate $P[X = x]$.

We will assume that

$$r > n \quad \text{and} \quad N - r > n,$$

so that we could have 0 to n black or red balls in our sample.

Then

$$\begin{aligned} & P[\text{exactly } x \text{ red balls out of } n \text{ selected}] \\ &= \frac{(\# \text{ ways to select } x \text{ out of } r \text{ balls}) \cdot (\# \text{ ways to select } n - x \text{ out of } N - r \text{ balls})}{\# \text{ ways to select } n \text{ out of } N \text{ balls}} \\ &= \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}} \end{aligned}$$

The Hypergeometric Distribution

9.1. Definition. Let $N, n, r \in \mathbb{N} \setminus \{0\}$, $r, n \leq N$, and $n < \min\{r, N - r\}$.

A random variable (X, f_X) with

$$X: S \rightarrow \Omega = \{0, \dots, n\}$$

and density function $f_X: \Omega \rightarrow \mathbb{R}$ given by

$$f_X(x) = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}} \quad (9.1)$$

is said to have a hypergeometric distribution with parameters N , n and r .

The Hypergeometric Identity

The hypergeometric distribution takes its name from the **hypergeometric identity**:

$$\binom{a+b}{r} = \sum_{k=0}^r \binom{a}{k} \binom{b}{r-k} = \sum_{i+j=r} \binom{a}{i} \binom{b}{j}. \quad (9.2)$$

To understand this identity, note that

$$\begin{aligned} \sum_{r=0}^{a+b} \binom{a+b}{r} x^r &= (1+x)^{a+b} = (1+x)^a (1+x)^b \\ &= \left(\sum_{i=0}^a \binom{a}{i} x^i \right) \left(\sum_{j=0}^b \binom{b}{j} x^j \right) \end{aligned}$$

The Hypergeometric Identity

Using the definition of the binomial coefficients (1.3), we see that $\binom{x}{i} = 0$ when $i > x$, so we may write

$$\begin{aligned}\sum_{r=0}^{\infty} \binom{a+b}{r} x^r &= \left(\sum_{i=0}^{\infty} \binom{a}{i} x^i \right) \left(\sum_{j=0}^{\infty} \binom{b}{j} x^j \right) \\ &= \sum_{r=0}^{\infty} \sum_{i+j=r} \binom{a}{i} \binom{b}{j} x^r,\end{aligned}$$

where we have used the Cauchy product of infinite series. Comparing term-by-term, (9.2) follows.

The hypergeometric identity is the main ingredient in showing that (9.1) actually defines a density function.

Non-independent Bernoulli Trials

Let us write

$$X = X_1 + X_2 + \cdots + X_n,$$

where each X_k is a Bernoulli random variable representing a single draw. Here “success” means drawing a red ball, yielding $X_k = 1$. If we draw a black ball on the k th draw, then $X_k = 0$.

We denote the probability of success by

$$p_k = P[X_k = 1]$$

Of course, the X_k are not independent - the result of each draw (X_k) influences the subsequent draws.

We therefore have to discuss the random vector (X_1, X_2, \dots, X_n) .

The Bernoulli Trials are Identical

To understand the distribution of the random vector (X_1, \dots, X_n) , we consider the sample space S : Suppose that we order the N balls in all conceivable ways, obtaining $N!$ permutations.

Effectively, we are drawing ***all balls*** out of the urn to obtain the sample space S , but only consider the events that include the ***first n*** balls to calculate the probabilities of our random vector.

The probability that $X_k = x$, where $x = 0$ or 1 , is then given by the number of elements in the sample space that have x in the k th position.

Since the sample space consists of ***all*** possible permutations of the N objects, we see that this probability does not depend on k . Therefore,

$$p_k = p_1 = \frac{r}{N}.$$

This shows that the Bernoulli trials are identical.

Expectation and Variance

We can calculate

$$\mathbb{E}[X_k] = 0 \cdot (1 - p_k) + 1 \cdot p_k = p_k = p_1 = \frac{r}{N}$$

so

$$\mathbb{E}[X] = \mathbb{E}[X_1 + \cdots + X_n] = \mathbb{E}[X_1] + \cdots + \mathbb{E}[X_n] = n \frac{r}{N}.$$

It is interesting to note that this expectation is the same as it would be if we were replacing the balls after drawing, i.e., if the number of red balls were determined by the binomial distribution.

In order to calculate the variance, we first generalize (8.1) to

$$\begin{aligned}\text{Var } X &= \text{Var}(X_1 + \cdots + X_n) \\ &= \text{Var } X_1 + \cdots + \text{Var } X_n + 2 \sum_{i < j} \text{Cov}(X_i, X_j).\end{aligned}$$

Variance and Covariance

We need to calculate

$$\text{Cov}(X_i, X_j) = E[X_i X_j] - E[X_i] E[X_j].$$

For this, we note that $X_i X_j$ is also a Bernoulli variable, since

$$X_i X_j = \begin{cases} 1 & \text{if } X_i = 1 \text{ and } X_j = 1, \\ 0 & \text{otherwise.} \end{cases}$$

Then

$$E[X_i X_j] = p_{ij} := P[X_i = 1 \text{ and } X_j = 1].$$

Variance and Covariance

As in the previous argument, the probability that $X_i = x$ and $X_j = y$ for $i \neq j$ and $x, y \in \{0, 1\}$ is given by the number of permutations among all $N!$ elements of the sample space that have x in the i th and y in the j th position. Again it is clear that this number is independent of i and j , so

$$p_{ij} = p_{12} = P[X_1 = 1 \text{ and } X_2 = 1] = \frac{r}{N} \cdot \frac{r-1}{N-1}.$$

Note that for $i = j$ we have

$$p_{ii} = p_{11} = P[X_1 = 1 \text{ and } X_1 = 1] = \frac{r}{N}.$$

Hence,

$$\text{Var } X_i = \frac{r}{N} \left(1 - \frac{r}{N}\right), \quad \text{Cov}(X_i, X_j) = -\frac{1}{N} \cdot \frac{r(N-r)}{N(N-1)}.$$

Approximating the Hypergeometric Distribution

Since there are $\binom{n}{2} = n(n - 1)/2$ pairs (i, j) with $i < j$, an easy calculation (do it yourself!) now gives

$$\text{Var } X = n \frac{r}{N} \frac{N - r}{N} \frac{N - n}{N - 1}$$

This expression is similar to that for the binomial distribution; if we were replacing the balls after drawing, we would have

$$p = \frac{r}{N}, \quad q = \frac{N - r}{N}$$

and since the variance of the binomial distribution is npq , we see that the expression above differs by

$$\frac{N - n}{N - 1}.$$

In fact, the binomial distribution may be used to approximate the hypergeometric distribution if the **sampling fraction** n/N is small (less than 0.05).

Approximating the Hypergeometric Distribution

9.2. Example. A production lot of 200 units has 8 defectives. A random sample of 10 units is selected, and we want to find the probability that the random sample will contain exactly one defective.

The true probability is

$$P[X = 1] = \frac{\binom{r}{x} \binom{N-r}{n-x}}{\binom{N}{n}} = \frac{\binom{8}{1} \binom{192}{9}}{\binom{200}{10}} = 0.288.$$

We note that the sampling fraction is $n/N = 10/200 = 0.05$, so we can use the binomial approximation.

Then $p = r/N = 8/200 = 0.04$ and

$$P[X = 1] \approx \binom{10}{1} (0.04)^1 (0.96)^9 = 0.277.$$

Transformation of Random Variables and Reliability

Transformation of Variables

The following theorem allows us to perform transformations of random variables and obtain the densities of the transformed variables.

10.1. Theorem. Let $(\mathbf{X}, f_{\mathbf{X}})$ be a continuous multivariate random variable and let $\varphi: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a differentiable, bijective map with inverse φ^{-1} . Then $\mathbf{Y} = \varphi \circ \mathbf{X}$ is a continuous multivariate random variable with density

$$f_{\mathbf{Y}}(y) = f_{\mathbf{X}} \circ \varphi^{-1}(y) \cdot |\det D\varphi^{-1}(y)|,$$

where $D\varphi^{-1}$ is the Jacobian of φ^{-1} .

We will not prove this theorem, which is based on the substitution rule for multivariable integrals.

Transformation of Variables

We can use transformation of bivariate random variables to obtain densities of sums and products of random variables, as the following example shows:

10.2. Lemma. Let $((X, Y), f_{XY})$ be a continuous bivariate random variable. Let $U = X/Y$. Then the density f_U of U is given by

$$f_U(u) = \int_{-\infty}^{\infty} f_{XY}(uv, v) \cdot |v| dv.$$

Proof.

Consider the transformation $\varphi: (X, Y) \mapsto (U, V)$ where

$$\varphi(x, y) = \begin{pmatrix} x/y \\ y \end{pmatrix}.$$

Then

$$\varphi^{-1}(u, v) = \begin{pmatrix} uv \\ v \end{pmatrix}.$$

Transformation of Variables

Proof.

We calculate

$$D\varphi^{-1}(u, v) = \begin{pmatrix} \frac{\partial x}{\partial u} & \frac{\partial x}{\partial v} \\ \frac{\partial y}{\partial u} & \frac{\partial y}{\partial v} \end{pmatrix} = \begin{pmatrix} v & u \\ 0 & 1 \end{pmatrix}$$

so

$$|\det D\varphi^{-1}(u, v)| = |v|.$$

Then

$$f_{UV}(u, v) = f_{XY}(uv, v)|v|.$$

The marginal density f_U is given by

$$f_U(u) = \int_{-\infty}^{\infty} f_{UV}(u, v) dv = \int_{-\infty}^{\infty} f_{XY}(uv, v) \cdot |v| dv.$$

□

The Chi Random Variable

Consider the following problem: a point $z = (z_1, \dots, z_n)$ in \mathbb{R}^n is randomly selected in such a way that every coordinate value z_i , $i = 1, \dots, n$, is determined independently of the other coordinates by a random variable Z_i . Suppose that each Z_i follows a standard normal distribution.

We are interested in the distribution function of the random variable

$$\chi_n := \sqrt{\sum_{i=1}^n Z_i^2}$$

which describes the distance of the selected point from the origin. For instance, while the expected value of each coordinate is $E[Z_i] = 0$, we do not know the expected distance from the origin, $E[\chi_n]$.

We say that χ_n is a **chi random variable** and that it follows a **chi distribution with n degrees of freedom**.

The Chi Distribution

To find the density f_{χ_n} , we consider the cumulative distribution function F_{χ_n} ,

$$F_{\chi_n}(y) = P[\chi_n \leq y].$$

Clearly, $F_{\chi_n}(y) = 0$ for $y < 0$. For $y \geq 0$,

$$\begin{aligned} F_{\chi_n}(y) &= P[\chi_n \leq y] = P[\chi_n^2 \leq y^2] = P\left[\sum_{k=1}^n Z_k^2 \leq y^2\right] \\ &= \int_{\sum_{k=1}^n z_k^2 \leq y^2} f_{Z_1 \dots Z_n}(z_1, \dots, z_n) dz_1 \dots dz_n \end{aligned}$$

Note that the n independent standard normal variables Z_1, \dots, Z_n have joint density

$$f_{Z_1 \dots Z_n}(z_1, \dots, z_n) = \frac{1}{(2\pi)^{n/2}} e^{-\sum_{k=1}^n z_k^2/2}.$$

The Chi Distribution

We hence obtain

$$F_{\chi_n}(y) = \int_{\sum_{k=1}^n z_k^2 \leq y^2} (2\pi)^{-n/2} e^{-\sum_{k=1}^n z_k^2/2} dz_1 \dots dz_n.$$

It becomes convenient to introduce polar coordinates $(r, \theta_1, \dots, \theta_{n-1})$ with $r > 0$, $0 < \theta_{n-1} < 2\pi$ and $0 < \theta_k < \pi$ for $k = 1, \dots, n-2$ as follows:

$$x_1 = r \cos \theta_1$$

$$x_2 = r \sin \theta_1 \cos \theta_2$$

$$x_3 = r \sin \theta_1 \sin \theta_2 \cos \theta_3$$

$$\vdots$$

$$x_{n-1} = r \sin \theta_1 \sin \theta_2 \dots \sin \theta_{n-2} \cos \theta_{n-1}$$

$$x_n = r \sin \theta_1 \sin \theta_2 \dots \sin \theta_{n-2} \sin \theta_{n-1}.$$

The Chi Distribution

The integral becomes

$$\begin{aligned} F_{\chi_n}(y) &= \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} \int_0^y (2\pi)^{-n/2} e^{-r^2/2} r^{n-1} \\ &\quad \times D(\theta_1, \dots, \theta_{n-1}) dr d\theta_1 \dots d\theta_{n-2} d\theta_{n-1} \end{aligned}$$

where $D(\theta_1, \dots, \theta_{n-1})$ is independent of r . Writing

$$C_n = (2\pi)^{-n/2} \int_0^{2\pi} \int_{-\pi/2}^{\pi/2} \cdots \int_{-\pi/2}^{\pi/2} D(\theta_1, \dots, \theta_{n-1}) d\theta_1 \dots d\theta_{n-2} d\theta_{n-1}$$

we have

$$F_{\chi_n}(y) = C_n \int_0^y e^{-r^2/2} r^{n-1} dr.$$

The Chi Distribution

We determine C_n from

$$1 = \lim_{y \rightarrow \infty} F_{\chi_n}(y) = C_n \int_0^{\infty} e^{-r^2/2} r^{n-1} dr = C_n \Gamma\left(\frac{n}{2}\right) 2^{n/2-1},$$

where we have substituted $\rho = r^2/2$ in the integral to obtain the gamma function. It follows that

$$F_{\chi_n}(y) = \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2-1}} \int_0^y e^{-r^2/2} r^{n-1} dr.$$

and the density of χ_n is given by

$$f_{\chi_n}(y) = F'_{\chi_n}(y) = \frac{2}{2^{n/2} \Gamma\left(\frac{n}{2}\right)} y^{n-1} e^{-y^2/2}. \quad (10.1)$$

for $y \geq 0$ (and $f_{\chi_n}(y) = 0$ for $y < 0$).

The Chi-Squared Distribution

In statistics, we will be particularly interested in the **chi-squared random variable** with n degrees of freedom,

$$\chi_n^2 = \sum_{i=1}^n Z_i^2. \quad (10.2)$$

where again Z_1, \dots, Z_n are independent standard normal random variables. Hence, a chi-squared random variable represents the sum of the squares of independent standard normal variables.

We obtain the density of χ_n^2 by again considering the cumulative distribution function: For $y \geq 0$,

$$\begin{aligned} F_{\chi_n^2}(y) &= P[\chi_n^2 \leq y] = P[-\sqrt{y} \leq Z_n \leq \sqrt{y}] \\ &= \frac{1}{\Gamma\left(\frac{n}{2}\right) 2^{n/2-1}} \int_0^{\sqrt{y}} e^{-r^2/2} r^{n-1} dr \end{aligned}$$

The Chi-Squared Distribution

Differentiating and applying the chain rule, we have

$$\begin{aligned}f_{\chi_n^2}(y) &= F'_{\chi_n^2}(y) = \frac{1}{\Gamma(\frac{n}{2}) 2^{n/2-1}} \frac{d}{dy} \int_0^{\sqrt{y}} e^{-r^2/2} r^{n-1} dr \\&= \frac{1}{2^{n/2} \Gamma(\frac{n}{2})} y^{n/2-1} e^{-y/2}.\end{aligned}$$

Now if $y < 0$,

$$F_{\chi_n^2}(y) = P[\chi_n^2 < y] \leq P[\chi_n^2 < 0] = 0,$$

so differentiation yields $f_{\chi_n^2}(y) = 0$ for $y < 0$.

The density $f_{\chi_n^2}$ is called a **chi-squared distribution**. We have already remarked that it is a gamma distribution with $\beta = 2$ and $\alpha = n/2$.

The Sum of Independent Chi-Squared Variables

Suppose we have two independent chi-squared random variables with m and n degrees of freedom, χ_m^2 and χ_n^2 . Then we can write

$$\chi_m^2 = \sum_{i=1}^m X_i^2, \quad \chi_n^2 = \sum_{j=1}^n Y_j^2$$

where the X_i and Y_j , $i = 1, \dots, m$, $j = 1, \dots, n$, are independent standard normal random variables. Now the sum

$$\chi_{m+n}^2 := \chi_m^2 + \chi_n^2 = \sum_{i=1}^m X_i^2 + \sum_{j=1}^n Y_j^2$$

is clearly the sum of $m + n$ squares of independent standard normal random variables. Therefore, it also follows a chi-squared distribution, but with $m + n$ degrees of freedom.

The Sum of Independent Chi-Squared Variables

We have the following general result:

10.3. Lemma. Let $\chi_{\gamma_1}^2, \dots, \chi_{\gamma_n}^2$ be n independent random variables following chi-squared distributions with $\gamma_1, \dots, \gamma_n$ degrees of freedom, respectively.

Then

$$\chi_{\alpha}^2 := \sum_{k=1}^n \chi_{\gamma_k}^2$$

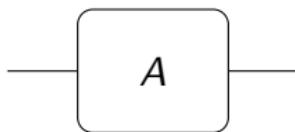
is a chi-squared random variable with $\alpha = \sum_{k=1}^n \gamma_k$ degrees of freedom.

Question. The sum of two chi-squared random variables is again a chi-squared random variable. What about the difference of two such variables?

- (1) The difference is also a chi-squared random variable.
- (2) The difference follows some other distribution.

A Black Box System

Consider a “black box” unit A :



We don't care what the unit A does or what it looks like inside. We simply assume that at time $t = 0$ the unit A is working. Then at any time $t > 0$, either

- ▶ A is working or
- ▶ A has failed.

When A fails, it fails completely and can not be repaired.

Failure Density

The time when A fails is random; we describe it by the continuous random variable T_A . The density of T_A is called the

failure density f_A .

The cumulative distribution function of T_A is denoted by F_A .

We note that

$$\begin{aligned} f_A(t) &= \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T \leq t + \Delta t]}{\Delta t} \\ &= \lim_{\Delta t \rightarrow 0} \frac{F_A(t + \Delta t) - F_A(t)}{\Delta t} \end{aligned} \tag{10.3}$$

Reliability Function

In practice, one often works with the

reliability function R_A .

The reliability function gives the probability that A is working at time $t \geq 0$.

By our assumption, $R_A(0) = 1$ and

$$\begin{aligned} R_A(t) &= 1 - P[\text{component } A \text{ fails before time } t] \\ &= 1 - \int_0^t f_A(s) \, ds \\ &= 1 - F_A(t). \end{aligned}$$

Hazard Rate

For practical purposes, an important quantity is the

hazard rate ϱ_A .

defined by

$$\varrho_A(t) := \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T \leq t + \Delta t \mid t \leq T]}{\Delta t}$$

(compare with (10.3)). We see that

$$\begin{aligned}\varrho_A(t) &= \lim_{\Delta t \rightarrow 0} \frac{P[t \leq T \leq t + \Delta t]}{P[T \geq t] \cdot \Delta t} \\ &= \frac{f_A(t)}{R_A(t)}.\end{aligned}$$

Question. The hazard rate is often directly observable in practice, while the failure density f is not. Why is this so?

Interpretation of the Hazard Rate

The hazard rate function ϱ can be interpreted qualitatively as follows:

- (i) If ϱ is decreasing over an interval, then as time goes by a failure is less likely to occur than it was earlier in the time interval. This happens in situations in which defective systems tend to fail early. As time goes by, the hazard rate for a well-made system decreases.
- (ii) A steady hazard rate is expected over the useful life span of a component. A failure tends to occur during this period due mainly to random factors.
- (iii) If ϱ is increasing over an interval, then as time goes by a failure is more likely to occur. This normally happens for systems that begin to fail primarily due to wear.

A typical component may exhibit all these behaviors over its lifetime, giving rise to a so-called ***bathtub curve***.

Finding the Reliability Function

Often one has information on ϱ , but not of the failure density f or reliability function R .

10.4. Theorem. Let X be a random variable with failure density f , reliability function R and hazard rate ϱ . Then

$$R(t) = e^{-\int_0^t \varrho(x) dx}.$$

Proof.

Since $R(x) = 1 - F(x)$ we have $R'(x) = -F'(x)$. Therefore,

$$\varrho(x) = \frac{f(x)}{R(x)} = \frac{F'(x)}{R(x)} = -\frac{R'(x)}{R(x)}$$

so

$$R'(x) = -\varrho(x)R(x).$$

Solving this equation with $R(0) = 1$ (why?), we obtain the result. □

The Weibull Density

10.5. Example. One hazard function in widespread use is the function

$$\varrho(t) = \alpha\beta t^{\beta-1}, \quad t > 0, \quad \alpha, \beta > 0$$

- ▶ If $\beta = 1$, the hazard rate is constant
- ▶ If $\beta > 1$, the hazard rate is increasing
- ▶ If $\beta < 1$, the hazard rate is decreasing

The reliability function is given by

$$R(t) = e^{-\int_0^t \alpha\beta x^{\beta-1} dx} = e^{-\alpha t^\beta}.$$

The failure density is given by

$$f(t) = \varrho(t)R(t) = \alpha\beta t^{\beta-1} e^{-\alpha t^\beta}.$$

This density is called the **Weibull density**, named after W. Weibull who introduced it in 1951.

Weibull Distribution

10.6. Definition. A random variable (X, f_X) is said to have a Weibull distribution with parameters α and β if its density is given by

$$f(x) = \begin{cases} \alpha\beta x^{\beta-1} e^{-\alpha x^\beta}, & x > 0, \\ 0, & \text{otherwise,} \end{cases} \quad \alpha, \beta > 0.$$



Waloddi Weibull (1887-1970) Abernethy, R. B.,
Waloddi Weibull- Historia, Extract from The New
Weibull Handbook.

10.7. Theorem. Let X be a Weibull random variable with parameters α and β . The mean and variance of X are given by

$$\mu = \alpha^{-1/\beta} \Gamma(1 + 1/\beta)$$

and

$$\sigma^2 = \alpha^{-2/\beta} \Gamma(1 + 2/\beta) - \mu^2.$$

 The Uniform Distribution

10.8. Example. Consider a uniform failure density of

$$f_X(x) = \begin{cases} 1 & 0 \leq x \leq 1, \\ 0 & \text{otherwise.} \end{cases}$$

Hence, failure is equally likely at any time between $x = 0$ and $x = 1$. In Mathematica, the uniform distribution is implemented as follows:

```
PDF[UniformDistribution[{0, 1}], x]
```

$$\begin{cases} 1 & 0 \leq x \leq 1 \\ 0 & \text{True} \end{cases}$$



The Uniform Distribution

Question. Use the Mathematica commands **SurvivalFunction** and **HazardFunction** to find the reliability function and the hazard rate for the uniform distribution on $[0, 1]$.

Systems in Series and Parallel Configurations

Components in multiple-component systems can be installed in the system in various ways. Many systems are arranged in “series” configuration, some are in “parallel” and others are combinations of the two designs.

10.9. Definition.

- (i) A system whose components are arranged in such a way that the system fails whenever any of its components fail is called a **series** system.
- (ii) A system whose components are arranged in such a way that the system fails only if all of its components fail is called a **parallel** system.

Reliability of Series and Parallel Systems

Assuming the components are independent of each other, the reliability of a series system with k components is given by

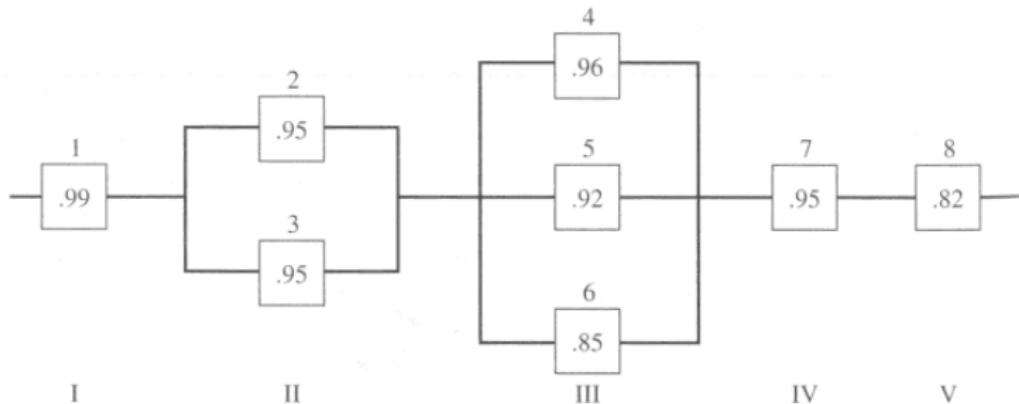
$$R_s(t) = \prod_{i=1}^k R_i(t),$$

where R_i is the reliability of the i th component. The reliability of a parallel system is given by

$$\begin{aligned} R_p(t) &= 1 - P[\text{all components fail before } t] \\ &= 1 - \prod_{i=1}^k (1 - R_i(t)). \end{aligned}$$

Reliability of Series and Parallel Systems

10.10. Example. Consider a system consisting of eight independent components, connected as shown below:



The numbers shown are the reliabilities $R(t_0)$ for fixed $t_0 > 0$. The reliability of the entire system is the product of assemblies I-V, working out to 0.77.

Samples and Data

Populations and Random Variables

A **population** is a large (possibly infinite) collection of individuals, objects or other instances which we want to describe probabilistically.

We assume that a population is described by a (scalar or multivariate) **random variable** in the following sense:

For each member of the population, the random variable takes on a given, deterministic, value.

The “randomness” of the random variable consists of the randomness of selecting an individual from the population.

Mathematically, we denote by $X = x$ the value of the random variable. Selecting a given member of the population and measuring X gives one instance of this random variable.

The probability density of X describes the likelihood of obtaining a value x within a given range.

Coin Flips

11.1. Example. Suppose we are flipping a coin with probability of heads p , $0 < p < 1$. The population might be

all flips of this coin conducted in the future

while the random variable X might be

$$X = \begin{cases} 1 & \text{coin turns heads up,} \\ 0 & \text{otherwise.} \end{cases}$$

Hence X follows a Bernoulli distribution with parameter p .

Each individual flip of the coin would represent an independent and identical copy of X .

We say that X describes the population, meaning that each member of the population gives an identical copy of X . The population size is indeterminate.

Student Height

11.2. Example. Suppose we are interested in the body height of the students of a certain university. Hence, our population might be described as

all students who were enrolled in the university in 2020.

The random variable X would be

the height (in cm) of the population.

It may well be that X is described by a normal distribution with certain mean and variance.

Each individual student would represent an independent and identical copy of X . The population size is possibly large, but is a well-defined number.

Again, the student height X describes the population, meaning that each student gives an identical copy of X .

Probability Theory vs. Statistics

Given a population and a random variable, probability theory and statistics are concerned with different questions:

Probability: The distribution of the random variable is fully known. What inferences can be drawn from the known information?

Statistics: The probability distribution is not known, but perhaps certain assumptions may be made. Data is gathered in order to make inferences on the distribution, e.g., its shape, expectation, variance etc.

In short: ***probability theory*** supposes one has complete knowledge of all parameters of a distribution, while ***statistics*** attempts to gain information on these parameters through experiments.

Probability Theory vs. Statistics

11.3. Examples.

- (i) When considering coin flips, probability theory might answer the question: if $p = 1/2$, what is the likelihood of obtaining more than 60 heads when performing 100 coin flips?

A statistical question would be: If one obtains more than 60 heads in 100 coin flips, what can be said about p ? Is there evidence that the coin is not fair? Can we give an interval $[p_0, p_1] \subset [0, 1]$ where we can be 90% sure that $p \in [p_0, p_1]$?

- (ii) Given a student population whose height follows a normal distribution with mean μ and variance σ^2 , probability theory would allow the calculation of the percentage of students whose height is above or below some threshold.

Statistics would attempt to gain information on μ and σ^2 by measuring the height of a certain number of students.

Random Sample (Mathematical Definition)

The remainder of this course is concerned with statistics, based on methods and techniques of probability theory.

The basis of all statistical approaches is a ***random sample***. The mathematical definition is straightforward:

11.4. Definition. A ***random sample of size n from the distribution of X*** is a collection of n independent random variables X_1, \dots, X_n , each with the same distribution as X .

We say that X_1, \dots, X_n are independent, identically distributed (i.i.d.) random variables.

Each population member is an identical copy of X . A random sample comprises an independent selection of these copies.

Random Samples in Practice

Heuristically, a random sample is a subset of a population whose members have been selected in such a way, that the selection of one member does not influence the selection of any other member.

This means that each member of a random sample has been selected completely at random from the entire population and there is no **bias** in the selection.

For example, in obtaining a random sample of coin flips, one might just flip the coin. This is straightforward.

But to obtain a random sample of students enrolled at a university, it is not sufficient to just walk into a classroom for a course in, e.g., mathematics and measure the height of all students found there.

Question. Why is this?

Sample Size

We will generally discuss a random sample of size n from a population.
How large should n be?

- ▶ The size n of a random sample should not be too small. However, a large population does not imply the need for a large random sample. In fact, given the need to make inferences to some specified degree of accuracy,

the required minimum size of n is absolute,
independent of the population size

- ▶ However, n should not be too large relative to the population size:

If n is greater than 5% of the population,
special care must be taken.

We will suppose that our sample sizes are always smaller than 5% of the population.

Sample Sizes That Are Too Large

11.5. Example. Suppose that you are interested in a population of 100 students (e.g., all graduate students in a certain school). You wish to know what the proportion of students with body height greater than 180 cm is.

Suppose that (unknown to you), 20 out of 100 students have this height or greater. Suppose you take a sample of 50 students.

Then, using the hypergeometric distribution, we can calculate that there is 10% chance that the random sample includes 13 or more students with this height. Your guess for the proportion of students is then

$$13/50 = 26\%.$$

However, in the remaining population, only $7/50 = 14\%$ actually have this height. The large sample has not only yielded a result that is different from the true proportion (that is to be expected in statistics), it has also ***perturbed the distribution of the remaining population.***

The 1936 US Presidential Election

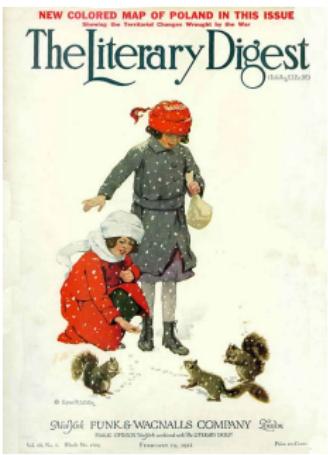
In 1936, Alfred Landon, Republican and governor of Kansas, was seeking to unseat the incumbent president of the United States, Franklin D. Roosevelt.

The magazine ***The Literary Digest***, attempted to predict the outcome of the election a month before by conducting one of the biggest polls ever: Based on magazine subscription data, club membership lists and telephone directories, it queried more than 10,000,000 individuals, sending them a mock ballot.

It received 2.4 million responses and concluded that Landon would win, 57% – 43%.

However, it turned out that Roosevelt won, 62% – 38%. What had gone wrong?

Literature: <https://www.math.upenn.edu/~deturck/m170/wk4/lecture/case1.html>



Cover of the vol. 68, issue 8 (number 1609) of
19 February 1936 edition of the *Literary Digest*. File:LiteraryDigest-19360219.jpg.
(2018, February 11). Wikimedia Commons, the free media repository.

Data

Suppose that we have obtained data from a random sample of size $n = 100$:

Data

```
{79, 141, 228, 3, 20, 14, 97, 194, 28, 56, 75, 37, 122, 27, 10,  
67, 23, 20, 103, 11, 92, 99, 64, 6, 118, 136, 682, 4, 70, 11,  
74, 40, 16, 114, 8, 149, 97, 7, 317, 346, 188, 149, 68, 150,  
88, 87, 155, 50, 26, 143, 126, 98, 153, 238, 30, 53, 132, 260,  
296, 25, 61, 87, 33, 51, 74, 111, 72, 178, 4, 67, 43, 229, 156,  
117, 104, 27, 23, 23, 186, 524, 107, 160, 41, 50, 352, 8, 153,  
142, 306, 320, 85, 44, 116, 39, 264, 360, 192, 142, 44, 29}
```

For most people this is just a “wall of numbers.” The first step in statistical analysis is to ***understand and visualize the data***. In this section, we will use the above data for our examples.

Percentiles and Quartiles

We can characterize data by using **percentiles**:

The x th percentile is defined as the value d_x of the data such that $x\%$ of the values of the data are less than or equal to d_x .

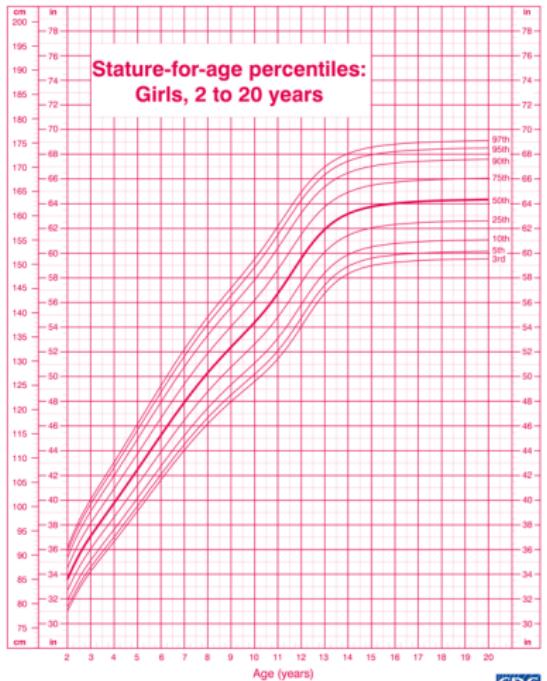
For instance, the 95th percentile is the datum such that 95% of the data is equal to or less than that value.

A special case are **quartiles**:

- ▶ 25% of the data are no greater than the **first quartile** q_1 ,
- ▶ 50% are no greater than the **second quartile** q_2 ,
- ▶ 75% are no greater than the **third quartile** q_3 .

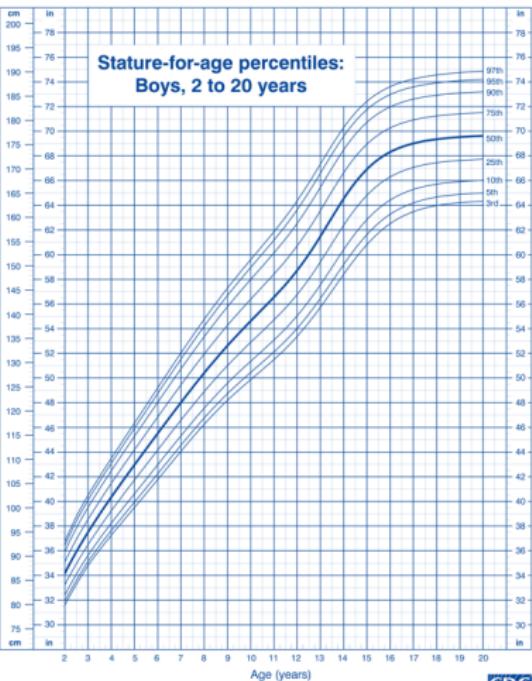
The second quartile is also known as the **median** of the data.
(You may compare with the notion of the median of a continuous distribution, introduced previously).

Percentile Growth Curves for US American Children



Published May 30, 2000

SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion.



Published May 30, 2000

SOURCE: Developed by the National Center for Health Statistics in collaboration with the National Center for Chronic Disease Prevention and Health Promotion (2000).

Calculating Quartiles

Suppose that our list of n data has been ordered from smallest to largest, so that

$$x_1 \leq x_2 \leq x_3 \leq \cdots \leq x_n.$$

Then the median is given by

$$q_2 = \begin{cases} x_{(n+1)/2} & \text{if } n \text{ is odd} \\ \frac{1}{2}(x_{n/2} + x_{n/2+1}) & \text{if } n \text{ is even} \end{cases}$$

Furthermore, the first quartile is defined as

- ▶ the median of the smallest $n/2$ elements if n is even.
- ▶ the average of the median of the smallest $(n - 1)/2$ elements and the median of the smallest $(n + 1)/2$ elements of the list if n is odd.

To calculate the third quartile, replace “smallest” with “largest” in the above definition.

★ Quartiles and Interquartile Range

Mathematica uses the definition of quartiles we have given here. Using the sample data shown before,

`Quartiles[Data]`

$$\left\{35, 87, \frac{299}{2}\right\}$$

The median (second quartile) is a measure of **location** of the data.

The difference between the third and first quartile is called the **interquartile range**,

$$\text{IQR} = q_3 - q_1,$$

and is a measure of **dispersion** of the data.

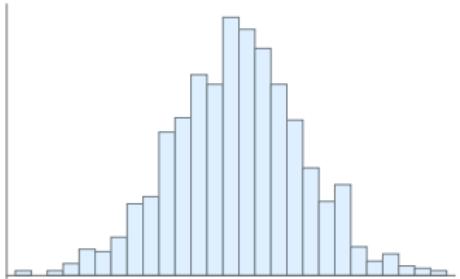
`InterquartileRange[Data]`

$$\frac{229}{2}$$

Histograms

A histogram is a (usually) vertical bar graph where each bar represents the (usually) the proportion or number of data in a given range.

The bars should show a rough silhouette of the underlying distribution's density function.



The histogram was first systematically introduced and analyzed by Karl Pearson.

Given data in a certain range, the first step is to select the number of categories, called **bins**, and correspondingly the width of each bin.

- ▶ **Too few bins:** The shape of the distribution can not be clearly distinguished, important features will be “smoothed out.”
- ▶ **Too many bins:** Individual bars are not supported by sufficiently many data points, spurious “features” may appear.

Number of Categories and Category Width

The traditional number of bins k is due to Sturges, which he proposed in 1926:

$$k = \lceil \log_2(n) \rceil + 1, \quad (11.1)$$

where the ***ceiling*** $\lceil x \rceil$ denotes the smallest integer greater than $x \in \mathbb{R}$.

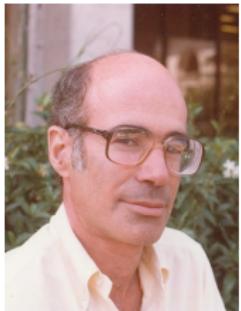
Sturges's rule is popular because it is simple and was based on one of the first serious analyses of this question. However, his derivation is flawed and the rule results in overly smoothed histograms for large n . Hence, various alternatives have been proposed.

We remark that the software Microsoft Excel uses the rule

$$k = \lceil \sqrt{n} \rceil.$$

Instead of the number k of categories, we can also fix the ***bin width h***.

The Freedman-Diaconis Rule



David A. Freedman (1938-2008)

File:David A Freedman (statistician) 1984.jpg. (2018, October 23). Wikimedia Commons, the free media repository.



Persi W. Diaconis (1945-)

File:Persi Diaconis 2010.jpg. (2014, April 18). Wikimedia Commons, the free media repository.

Among the various improvements that have been suggested, we will use the Freedman-Diaconis rule for the bin width h , which was presented in a publication in 1981.

The rule is designed to minimize the difference between the actual density of the distribution of the data and the height of the bars.

More precisely, if data of size n from the distribution of (X, f_X) is gathered on an interval I , then h should be chosen so that

$$\delta^2(h) = E \left[\int_I |H(x) - f_X(x)|^2 dx \right]$$

is minimized, where H is the normalized height of the histogram bars.

The Freedman-Diaconis Rule

Analysis shows that this is realized if the bin width is

$$h \sim \frac{1}{\sqrt[3]{n}} \quad \text{as } n \rightarrow \infty.$$

According to Freedman and Diaconis, numerical calculations show that

$$h = \frac{2 \cdot \text{IQR}}{\sqrt[3]{n}}$$

yields good results for the estimation of the true density f_X from the histogram.

Literature: Freedman, D., Diaconis, P. On the histogram as a density estimator: L_2 theory. *Z. Wahrscheinlichkeitstheorie verw. Gebiete* 57, 453–476 (1981).

<https://doi.org/10.1007/BF01025868>

Determining the Bin Widths

The **precision** of the data $\{x_1, \dots, x_n\}$ is the smallest decimal place of the values x_i .

The **sample range** is given by

$$\max_{1 \leq i \leq n} \{x_i\} - \min_{1 \leq i \leq n} \{x_i\}.$$

If the number of bins k has been determined (e.g., by Sturges's rule), then the bin width is calculated as

$$h = \frac{\max\{x_i\} - \min\{x_i\}}{k},$$

which should be rounded up to the precision of the data. If h is already at the precision of the data, one smallest decimal unit should be added to h .

If the bin width has been determined (e.g., by the Freedman-Diaconis rule), then nothing else needs to be done.

Binning the Data

Next, the actual bins need to be determined. Ideally, the bins should have the properties that

- ▶ The bins represent the data range well and do not go too far beyond it.
- ▶ Each datum should fall into exactly one bin.
- ▶ The bins should have the same width (in our approach here).

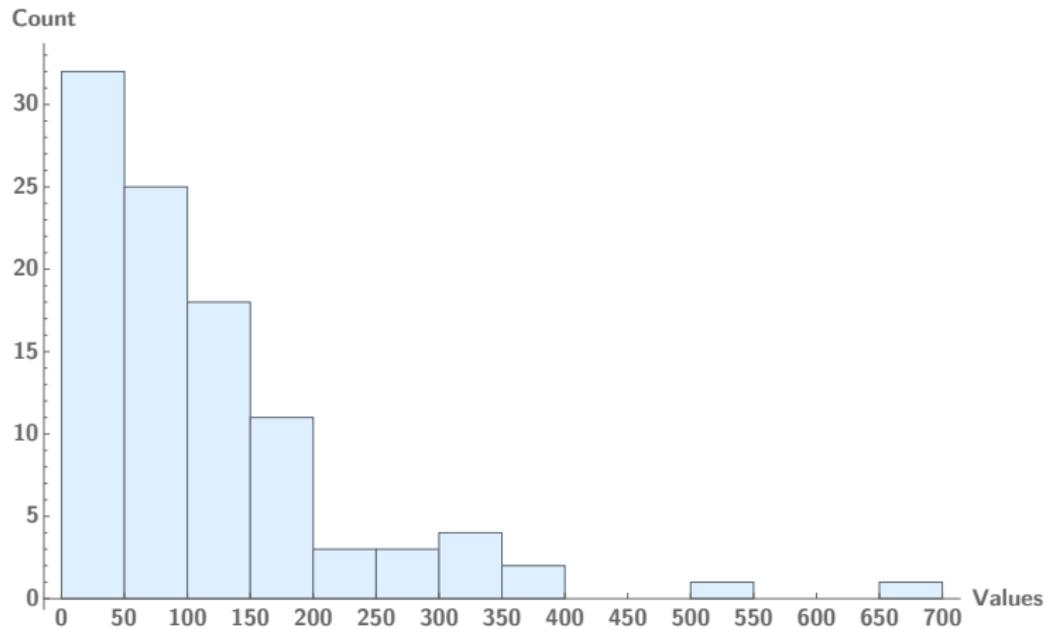
To achieve this, the ideal way is to take the smallest datum, subtract ***one-half of the smallest decimal of the data*** and then successively add the bin width to obtain the bins.

Since the bin boundaries are now at a higher precision than the data, no datum can lie on the boundary. The rounding up of the bin widths (if determined as above) will ensure that the data range is covered.

In practice, however, one often chooses “nice” values as bin boundaries.

Histogram (Freedman-Diaconis Bin Widths)

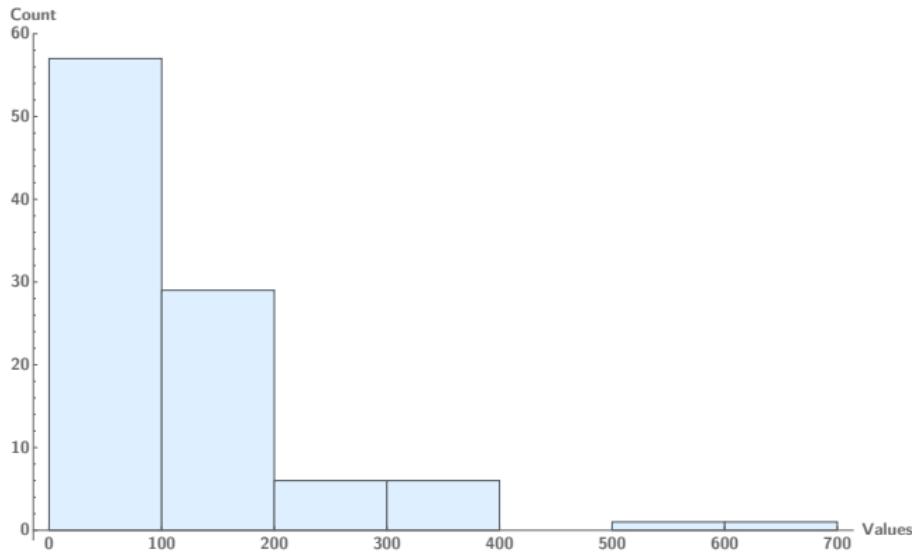
In our example, we have $\frac{2 \cdot \text{IQR}}{\sqrt[3]{n}} = 49.34$, which we round up to 50.



Mathematica: `Histogram[Data, "FreedmanDiaconis"]`

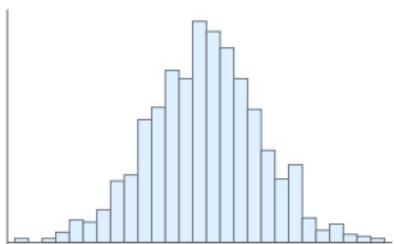
Histogram (Sturges's Rule Bin Number)

The data range is $682 - 3 = 679$ and Sturges's rule (based on 100 data) gives $k = 7$. We calculate $679/7 = 97$, which should be rounded up by one to $h = 98$.

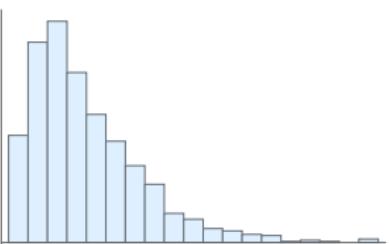


Mathematica: `Histogram[Data, "Sturges"]`

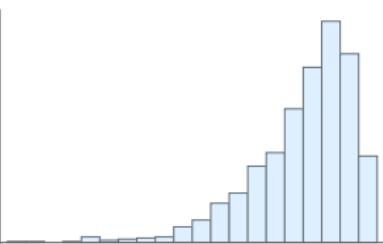
Describing a Histogram



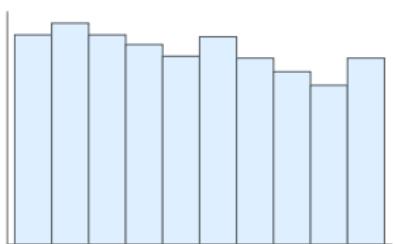
Symmetric,
unimodal



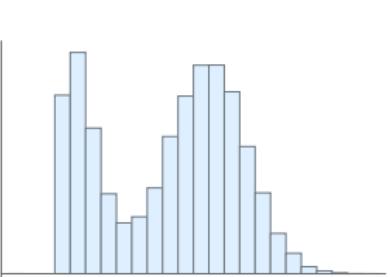
Positive skew,
unimodal



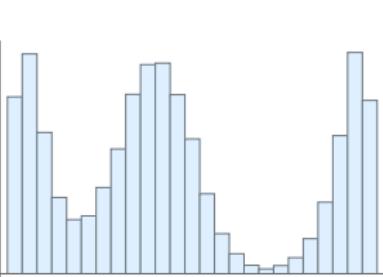
Negative skew,
unimodal



Symmetric,
no prominent mode



Bimodal



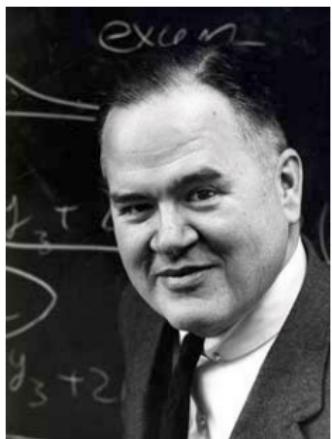
Multimodal

Stem-and-Leaf Diagrams

A ***stem-and-leaf diagram*** is a rough way to get an idea of the shape of the distribution of a random sample, while preserving some of its numeric information. It consists of labeled rows of numbers, where the label is called the stem and the other numbers are called leaves. This idea was introduced by Tukey in his famous book ***Exploratory data Analysis*** in 1977.

To construct a stem-and-leaf diagram from a random sample, follow these steps:

- (i) Choose a convenient number of leading decimal digits to serve as stems,
- (ii) label the rows using the stems,
- (iii) for each datum of the random sample, note down the digit following the stem in the corresponding row,
- (iv) turn the graph on its side to get an impression of its distribution.



John W. Tukey (1915-2000)
http://1stmouse.com/theJterm_software/

 Stem-and-Leaf Diagrams

We will continue to use the data of Slide 261.

The package `StatisticalPlots` includes a command for stem-and-leaf plots:

```
Needs["StatisticalPlots`"]
```

```
StemLeafPlot[Floor[Data, 10], IncludeEmptyStems → True]
```

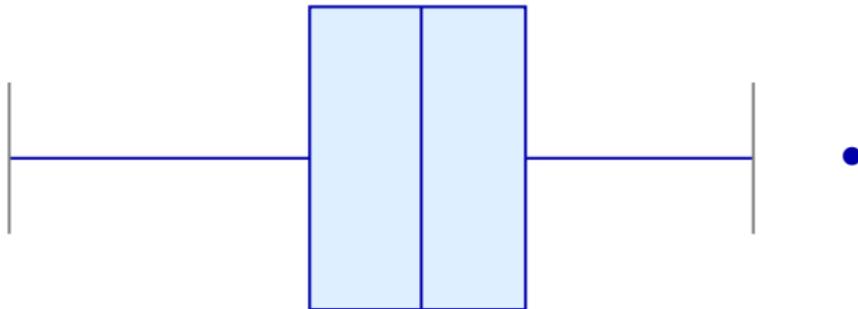
Stem	Leaves
0	000000011112222222233344445555666677777888899999
1	00011111223344444455555678899
2	223669
3	012456
4	
5	2
6	8

Stem units: 100

Box-and-Whisker Plots (Boxplots)

A **boxplot** is a representation of data that is useful for checking for symmetry or skew and, in general, deviation of the data from that expected of a normal distribution. Boxplots were also introduced by Tukey in his 1977 book.

This is their general appearance:



Construction of Boxplots

A boxplot is drawn on an abscissa scale of values corresponding to the data. Often, the abscissa scale is not shown.

The central box has a center line, located at the median q_2 , while the left and right sides of the box are located at the first and third quartiles q_1 and q_3 , respectively.

We define the **inner fences** f_1 and f_2 using the interquartile range as follows:

$$f_1 = q_1 - \frac{3}{2} \text{ IQR}, \quad f_3 = q_3 + \frac{3}{2} \text{ IQR}.$$

The “whiskers” (lines extending to the left and right of the box) end at the **adjacent values**

$$a_1 = \min\{x_k : x_k \geq f_1\}, \quad a_3 = \max\{x_k : x_k \leq f_3\}.$$

Construction of Boxplots

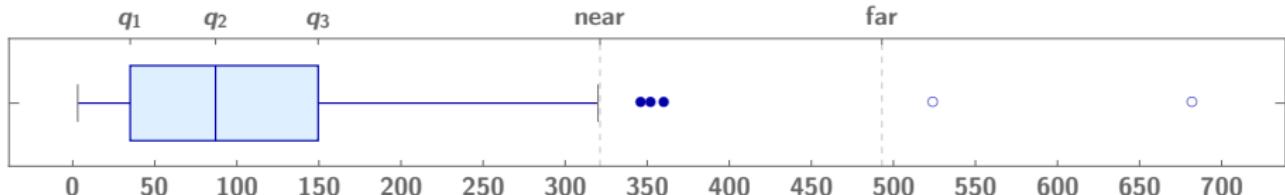
We define the **outer fences**

$$F_1 = q_1 - 3 \text{IQR},$$

$$F_3 = q_3 + 3 \text{IQR}.$$

Measurements x_k that lie outside the inner fences but inside the outer fences are called **near outliers**. Those outside the outer fences are known as **far outliers**.

A boxplot generated from the example data of Slide 261 is shown below:



Interpreting Boxplots

If data is obtained from a normal distribution, one would expect to see

- ▶ a symmetric median line in the middle of the box;
- ▶ equally long whiskers;
- ▶ very few near outliers and no far outliers.

A rule of thumb states that:

Of 1000 random samples of a normally distributed population, it can be expected that 7 will be outliers.

Data points lying between the inner and outer fences are called ***near outliers***, those lying outside the outer fences are called ***far outliers***. Far outliers are unusual if (and only if!) an approximately bell-shaped distribution of the random variable X of the population is expected. In this case, their origin should be investigated.

- ▶ If the outlier seems to be the result of an error in measurement or data collecting, it may be discarded from the data.
- ▶ If the outlier seems to be the result of a random measurement, it is recommended that statistics are reported twice: with the outlier included ***and*** without the outlier.

Interpreting Boxplots

A set of 10 data yields the following boxplot:



Which of the following sentences is the **most appropriate** conclusion?

- 1) There is no strong evidence that the data does not follow a normal distribution.
- 2) There is no strong evidence that the data follows a normal distribution.
- 3) There is strong evidence that the data does not follow a normal distribution.
- 4) There is strong evidence that the data follows a normal distribution.

Parameter Estimation

Statistics and Estimation

A random variable that is derived from a random sample X_1, \dots, X_n of a population is said to be **statistic**. Examples include

- ▶ any of the sample quartiles q_1, q_2, q_3 ,
- ▶ the sample maximum $\max\{X_1, \dots, X_n\}$,
- ▶ the sample mean

$$\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i.$$

We would like to use a given sample statistic to **estimate** a population parameter.

For example, the sample mean \bar{X} can be used to estimate the population mean μ .

Any statistic that is used in such way is then called an **estimator** and the value of the statistic a **point estimate**.

Bias and Mean Square Error

We would like an estimator to have the following properties:

- The expected value of $\hat{\theta}$ should be equal to θ ,
- $\hat{\theta}$ should have small variance for large sample sizes.

This motivates the following definition:

12.1. Definition. The difference

$$\theta - E[\hat{\theta}]$$

is called the **bias** of an estimator $\hat{\theta}$ for a population parameter θ . If $E[\hat{\theta}] = \theta$, we say that $\hat{\theta}$ is **unbiased**.

The **mean square error** of $\hat{\theta}$ is defined as

$$MSE(\hat{\theta}) := E[(\hat{\theta} - \theta)^2].$$

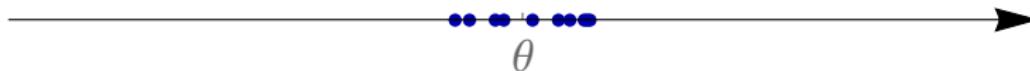
Quality of Estimators

The mean square error measures the overall quality of an estimator. We can write

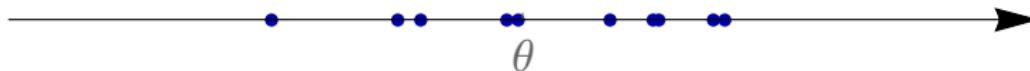
$$\begin{aligned}\text{MSE}(\hat{\theta}) &= E[(\hat{\theta} - E[\hat{\theta}])^2] + (\theta - E(\hat{\theta}))^2 \\ &= \text{Var } \hat{\theta} + (\text{bias})^2.\end{aligned}$$

Hence variance can be just as important as bias for an estimator. In general, unbiased estimators are preferred but sometimes biased estimators are used.

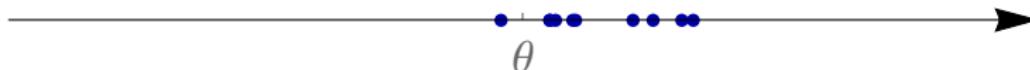
Simulation of 10 Estimates of θ



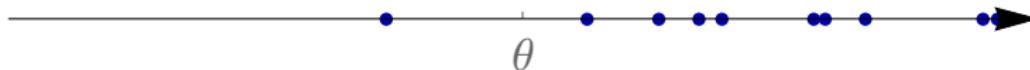
Unbiased, small variance



Unbiased, large variance



Biased, small variance



Biased, large variance

Sample Mean

12.2. Theorem. Let X_1, \dots, X_n be a random sample of size n from a distribution with mean μ . The sample mean \bar{X} is an unbiased estimator for μ .

Proof.

We simply insert the definition of the sample mean and use the properties of the expectation:

$$\begin{aligned} E[\bar{X}] &= E[(X_1 + \dots + X_n)/n] = \frac{1}{n} E[X_1 + \dots + X_n] \\ &= \frac{1}{n}(E[X_1] + \dots + E[X_n]) = \frac{n\mu}{n} = \mu. \end{aligned}$$

□

12.3. Theorem. Let \bar{X} be the sample mean of a random sample of size n from a distribution with mean μ and variance σ^2 . Then

$$\text{Var } \bar{X} = E[(\bar{X} - \mu)^2] = \frac{\sigma^2}{n}.$$

Sample Variance

Proof.

We simply insert the definition of the sample mean and use the properties of the variance:

$$\begin{aligned}\text{Var } \bar{X} &= \text{Var}((X_1 + \cdots + X_n)/n) = \frac{1}{n^2} \text{Var}(X_1 + \cdots + X_n) \\ &= \frac{1}{n^2}(\text{Var } X_1 + \cdots + \text{Var } X_n) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}.\end{aligned}$$

□

Thus \bar{X} is both unbiased and has a variance that decreases with large n ; it is a “nice” estimator, since we can make the mean square error $\text{MSE } \bar{X}$ as small as desired by taking n large enough.

12.4. Definition. The standard deviation of \bar{X} is given by $\sqrt{\text{Var } \bar{X}} = \sigma/\sqrt{n}$ and is called the **standard error of the mean**.

Medieval Standard Units of Measurement



Determination of the rood and the foot in Frankfurt, in J. Köbel, *Geometrie. Von künstlichem Feldmessen und Absehen*, published in Frankfurt, 1570. Image via Wikimedia, http://commons.wikimedia.org/wiki/File:Determination_of_the_rute_and_the Jeet_in_Frankfurt.png

Medieval Standard Units of Measurement

The picture on the previous slide is taken from the book “Geometrei. Von künstlichem Feldmessen und Absehen”, published in Frankfurt at the beginning of the 16th century. An edition from 1570 is available online at <http://books.google.de/books?id=80JSAAAAcAAJ&pg=PA1>.

The book describes the recommended method for obtaining a measurement of “1 foot” (although it doesn’t actually use the term):

*Sixteen men, small and large, as they freely leave the church one after the other, are each to put in front of the other a shoe. This same length is and shall be a right and proper measuring rood.
[...] Using a compass, this same measured rood is to be divided and distinguished into sixteen equal parts and shall forthwith be accepted and recognized as a right measuring rood for use in the field.*

Standard Error of the Mean and Sample Variance

Question. If the standard deviation of the size of shoes or feet in a population is σ , what is the standard deviation of 1/16 rood?

- (1) σ
- (2) $\sigma/2$
- (3) $\sigma/4$
- (4) $\sigma/16$

(In Japan, $\mu = 24.9$ cm and $\sigma = 1.05$ cm.)

The Method of Moments

General problem: How to find an estimator for a parameter of a distribution?

The **method of moments** was developed by Chebyshev and Pearson towards the end of the 19th century.

It is based on the fact that, given a random sample X_1, \dots, X_n of a random variable X , for any integer $k \geq 1$,

$$\widehat{E[X^k]} = \frac{1}{n} \sum_{i=1}^n X_i^k$$

is an unbiased estimator for the k th moment of X . (The proof is completely the same as for the sample mean.)

In other words, we have good estimators for the moments of a random variable.

The Method of Moments

The idea is now to express a parameter in terms of moments and then simply insert the estimators for these moments to obtain an estimator for the parameter.

Advantage: This is a simple method to obtain a basic estimator for a parameter.

Disadvantage: The estimators may not be unbiased and may yield non-sensical results in some cases.

For example, the variance of a random variable is $\text{Var}[X] = E[X^2] - E[X]^2$, so we can set

$$\begin{aligned}\widehat{\sigma^2} &= \widehat{E[X^2]} - \widehat{E[X]}^2 = \frac{1}{n} \sum_{i=1}^n X_i^2 - \bar{X}^2 \\ &= \frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2\end{aligned}$$

Estimator for the Variance

However, this estimator is not unbiased:

$$\begin{aligned} & \mathbb{E} \left[\sum_{k=1}^n (X_k - \bar{X})^2 \right] \\ &= \mathbb{E} \left[\sum_{k=1}^n (X_k - \mu + \mu - \bar{X})^2 \right] \\ &= \mathbb{E} \left[\sum_{k=1}^n (X_k - \mu)^2 - 2(\bar{X} - \mu) \sum_{k=1}^n (X_k - \mu) + n(\mu - \bar{X})^2 \right] \\ &= \mathbb{E} \left[\sum_{k=1}^n (X_k - \mu)^2 - 2(\bar{X} - \mu) \left(\left(\sum_{k=1}^n X_k \right) - n\mu \right) + n(\mu - \bar{X})^2 \right] \\ &= \mathbb{E} \left[\sum_{k=1}^n (X_k - \mu)^2 - 2(\bar{X} - \mu)(n\bar{X} - n\mu) + n(\mu - \bar{X})^2 \right]. \end{aligned}$$

Sample Variance

Simplifying, we have

$$\begin{aligned} E\left[\sum_{k=1}^n (X_k - \bar{X})^2\right] &= E\left[\sum_{k=1}^n (X_k - \mu)^2 - n(\bar{X} - \mu)^2\right] \\ &= \left(\sum_{k=1}^n E[(X_k - \mu)^2] - n E[(\bar{X} - \mu)^2]\right). \end{aligned}$$

We now use that

$$\begin{aligned} E[(X_k - \mu)^2] &= \text{Var}[X_k] = \sigma^2, \\ E[(\bar{X} - \mu)^2] &= \text{Var}[\bar{X}] = \sigma^2/n. \end{aligned}$$

Sample Variance

Then

$$E\left[\sum_{k=1}^n (X_k - \bar{X})^2\right] = \left(\sum_{k=1}^n \sigma^2 - n\frac{\sigma^2}{n}\right) = (n-1)\sigma^2.$$

It follows that the estimator $\widehat{\sigma^2}$ obtained by the method of moments is biased:

$$E\left[\frac{1}{n} \sum_{k=1}^n (X_k - \bar{X})^2\right] = \frac{n-1}{n} \sigma^2,$$

therefore this estimator would tend to underestimate the true variance.

Instead, we will work with the unbiased **sample variance**

$$S^2 := \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2.$$

Method of Maximum Likelihood

Fisher developed the following approach to finding estimators. Initial ideas go back to Gauß who used similar approaches on certain problems.

Given a set of observations x_1, \dots, x_n from a random variable X , with parameter θ one finds the value of θ most likely to have produced these observations. This value becomes the estimate $\hat{\theta}$.

In other words, we express the probability of obtaining x_1, \dots, x_n as a function of the parameter θ and then determine the value of θ that maximizes this probability.

Method of Maximum Likelihood

Let X_θ be a random variable with parameter θ and density f_{X_θ} . Given a random sample (X_1, \dots, X_n) that yielded values (x_1, \dots, x_n) we define the **likelihood function** L by

$$L(\theta) = \prod_{i=1}^n f_{X_\theta}(x_i).$$

If X_θ is a discrete random variable, then $L(\theta)$ is just the probability of obtaining the observed measurements:

$$P[X_1 = x_1 \text{ and } \dots \text{ and } X_n = x_n] = \prod_{i=1}^n P[X_i = x_i] = \prod_{i=1}^n f_{X_\theta}(x_i)$$

If X_θ is continuous, it represents the probability density.

We then maximize $L(\theta)$. The location of the maximum is then chosen to be the estimator $\hat{\theta}$.

Estimating the Poisson Parameter

12.5. Example. Suppose it is known that X follows a Poisson distribution with parameter k and we wish to estimate k .

The density for X is given by $f_k(x) = \frac{e^{-k} k^x}{x!}$, $x \in \mathbb{N}$. Given a random sample X_1, \dots, X_n the likelihood function is

$$L(k) = \prod_{i=1}^n f_k(x_i) = e^{-nk} \frac{k^{\sum x_i}}{\prod x_i!}.$$

To simplify our calculations, we take the logarithm:

$$\ln L(k) = -nk + \ln k \sum_{i=1}^n x_i - \ln \prod x_i!.$$

Maximizing $\ln L(k)$ will also maximize $L(k)$.

Estimating the Poisson Parameter

We take the first derivative and set it equal to zero:

$$\frac{d \ln L(k)}{dk} = -n + \frac{1}{k} \sum_{i=1}^n x_i = 0$$

so we find

$$\hat{k} = \bar{x}.$$

This is not unexpected, since $\mu = k$ for the Poisson distribution and \bar{X} is a good estimator for μ . In this case the maximum-likelihood estimator coincides with the method-of-moments estimator.

This method can be generalized to finding estimators for multiple parameters. In that case, the maximum of $L(\theta_1, \dots, \theta_j)$ is found with respect to all j variables.



Estimators with Mathematica

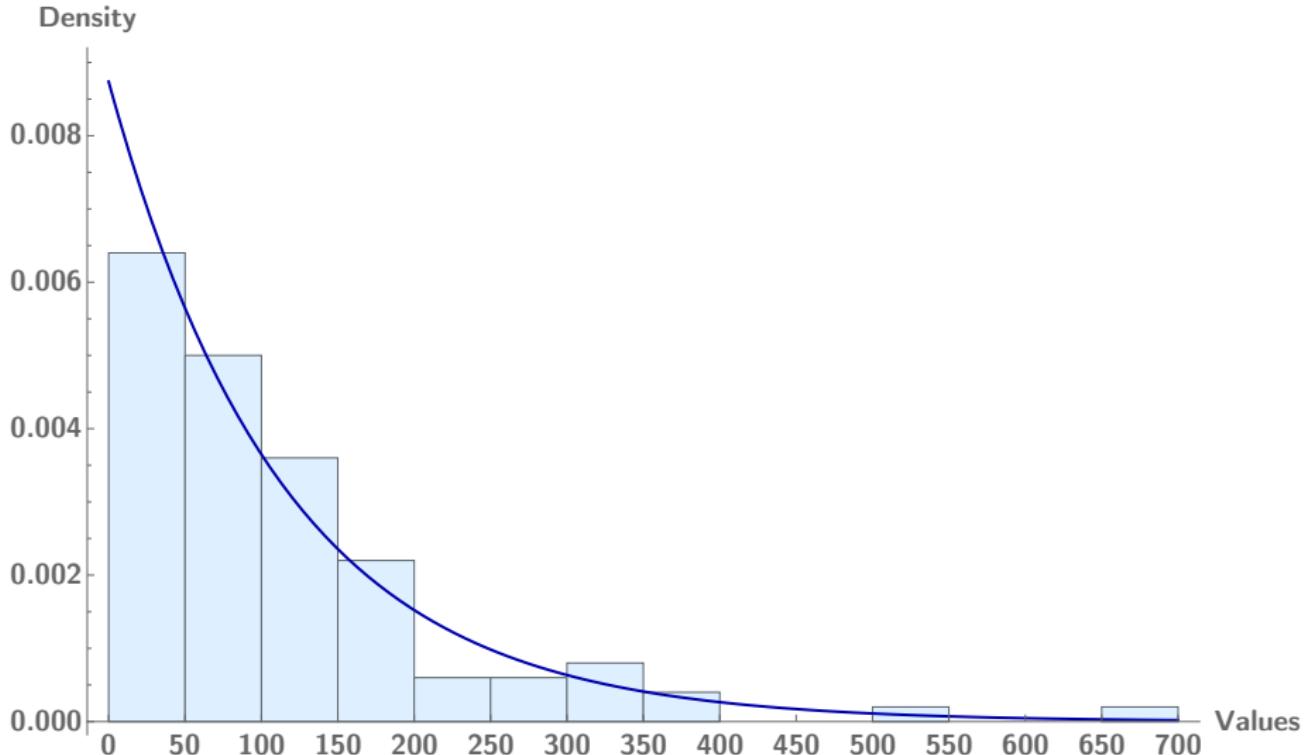
Both the method of moments and the method of maximum likelihood (as well as other methods) are available. Using the data from the previous section,

```
maxlike = FindDistributionParameters[Data,  
    ExponentialDistribution[\mathcal{B}],  
    ParameterEstimator -> "MaximumLikelihood"]  
  
\{\mathcal{B} \rightarrow 0.0087382\}
```

```
mom = FindDistributionParameters[Data,  
    ExponentialDistribution[\mathcal{B}],  
    ParameterEstimator -> "MethodOfMoments"]  
  
\{\mathcal{B} \rightarrow 0.0087382\}
```



Estimators with Mathematica



Interval Estimation

Distribution of the Sample Mean

Wanted: More precise information on estimated parameters.

Our goal now is to gain more precise information on the value of an estimated parameter. What we have obtained so far are point estimates, but we do not yet know how close such an estimate is to the actual value of the parameter.

Needed: the distribution of the sample statistic.

As a first example, let us consider the sample mean:

13.1. **Theorem.** Let X_1, \dots, X_n be a random sample of size n **from a normal distribution** with mean μ and variance σ^2 .

Then \bar{X} is normally distributed with mean μ and variance σ^2/n .

13.2. **Remark.** Even if the sample is taken from a non-normal distribution, if n is “sufficiently large”, then the distribution of \bar{X} will be close to normal due to the Central Limit Theorem 7.13.

Interval Estimation

13.3. Notation. We will often denote an interval of the form $[x - \varepsilon, x + \varepsilon]$ for $x \in \mathbb{R}$, $\varepsilon > 0$ by $x \pm \varepsilon$. In fact, we define

$$y = x \pm \varepsilon \quad \text{to mean} \quad y \in [x - \varepsilon, x + \varepsilon].$$

We would like to make statements such as “based on the results of a sample, we are 90% certain that the mean of a population lies in $\bar{X} \pm L$.”

This is known as ***interval estimation*** and the resulting interval is called a ***confidence interval***.

Two-Sided Confidence Intervals

13.4. Definition. Let $0 \leq \alpha \leq 1$. A $100(1 - \alpha)\%$ (**two-sided confidence interval for a parameter θ**) is an interval $[L_1, L_2]$ such that

$$P[L_1 \leq \theta \leq L_2] = 1 - \alpha. \quad (13.1)$$

13.5. Remark. The equation (13.1) does not determine L_1 and L_2 uniquely; we will nearly always require **centered confidence intervals** with

$$P[\theta < L_1] = P[\theta > L_2] = \alpha/2.$$

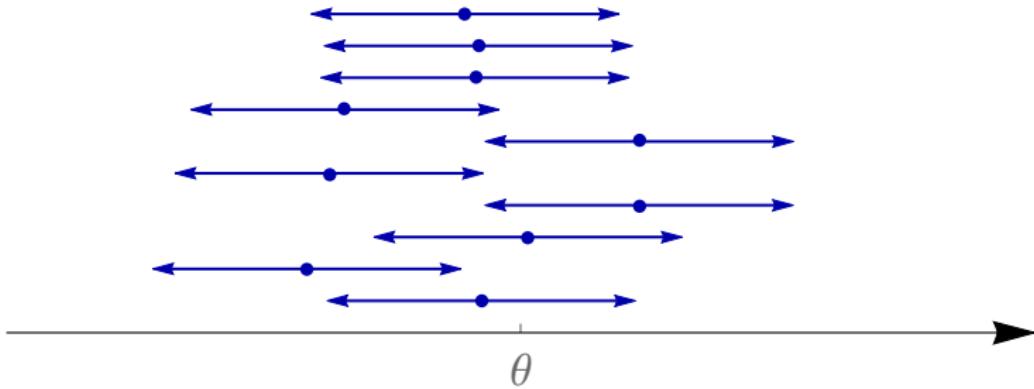
If the distribution of $\hat{\theta}$ is symmetric about θ , then

$$L_1 = \hat{\theta} - L, \quad L_2 = \hat{\theta} + L,$$

where L is a sample statistic and the interval is centered on $\hat{\theta}$, the point estimate for θ .

Random Intervals

13.6. Remark. It is important to note that in (13.1), the population parameter θ is not random, but that L_1 and L_2 are random. Hence, we may say that $[L_1, L_2]$ is a random interval.



Given θ , a random sample has a probability of $1 - \alpha$ of yielding sample statistics L_1 and L_2 such that $\theta \in [L_1, L_2]$.

Interval Estimation for the Mean (Variance Known)

Suppose that we have a random sample of size n from a normal population with **unknown mean** μ and **known variance** σ^2 .

A sample yields a point estimate \bar{X} for μ . We want to find $L = L(\alpha)$ such that we can state with $100(1 - \alpha)\%$ confidence that $\mu = \bar{X} \pm L$.

In particular, we would like to find a number L so that

$$P[\bar{X} - L \leq \mu \leq \bar{X} + L] = 1 - \alpha.$$

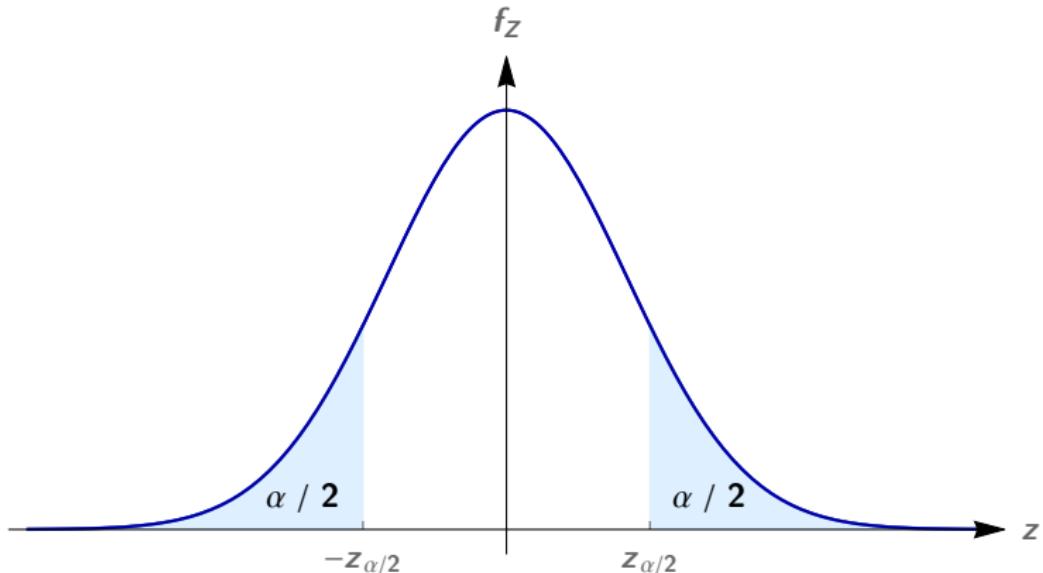
Note again that μ is not random, but rather a fixed but unknown parameter. However, the sample statistic \bar{X} is random and so is L .

It is crucial that we know the distribution of \bar{X} .

The Point $z_{\alpha/2}$

Given $\alpha \in [0, 1]$ we define $z_{\alpha/2} \in [0, \infty)$ by

$$\alpha/2 = P[Z \geq z_{\alpha/2}] = \frac{1}{\sqrt{2\pi}} \int_{z_{\alpha/2}}^{\infty} e^{-x^2/2} dx. \quad (13.2)$$



Interval Estimation for the Mean (Variance Known)

Fix $\alpha \in [0, 1]$. Then

$$1 - \alpha = P[\bar{X} - L \leq \mu \leq \bar{X} + L] = P\left[\frac{\bar{X} - \mu - L}{\sigma/\sqrt{n}} \leq 0 \leq \frac{\bar{X} - \mu + L}{\sigma/\sqrt{n}}\right]$$

By Theorem 13.1 the sample mean is normally distributed with mean μ and variance σ^2/n . Thus,

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}$$

follows a standard normal distribution, and so

$$\begin{aligned}1 - \alpha &= P\left[Z - \frac{L}{\sigma/\sqrt{n}} \leq 0 \leq Z + \frac{L}{\sigma/\sqrt{n}}\right] \\&= P\left[-\frac{L}{\sigma/\sqrt{n}} \leq Z \leq \frac{L}{\sigma/\sqrt{n}}\right] \\&= 2P\left[0 \leq Z \leq \frac{L}{\sigma/\sqrt{n}}\right] = 1 - 2P\left[\frac{L}{\sigma/\sqrt{n}} \leq Z < \infty\right]\end{aligned}$$

Confidence Interval for the Mean (Variance Known)

In this way we determine L as being the number such that

$$P\left[\frac{L}{\sigma/\sqrt{n}} \leq Z < \infty\right] = \alpha/2.$$

This is equivalent to writing

$$\frac{L}{\sigma/\sqrt{n}} = z_{\alpha/2} \quad \text{or} \quad L = \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}.$$

We have proved the following result:

13.7. Theorem. Let X_1, \dots, X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2 . A $100(1 - \alpha)\%$ confidence interval on μ is given by

$$\bar{X} \pm \frac{z_{\alpha/2} \cdot \sigma}{\sqrt{n}}.$$

Confidence Interval for the Mean (Variance Known)

13.8. Example. An article in the *Journal of Heat Transfer* describes a method of measuring the thermal conductivity of Armco iron. Using a temperature of 100°F and a power input of 550W , the following 10 measurements of thermal conductivity (in $\text{Btu}/(\text{hr ft }^{\circ}\text{F})$) were obtained:

$$\begin{array}{ccccc} 41.60 & 41.48 & 42.34 & 41.95 & 41.86 \\ 42.18 & 41.72 & 42.26 & 41.81 & 42.04 \end{array}$$

A point estimate of the mean thermal conductivity at 100°F and 550W is the sample mean,

$$\bar{x} = 41.92 \text{ Btu}/(\text{hr ft }^{\circ}\text{F}).$$

Suppose we know that the standard deviation of the thermal conductivity under the given conditions is $\sigma = 0.10 \text{ Btu}/(\text{hr ft }^{\circ}\text{F})$. A 95% confidence interval ($\alpha = 0.05$) on the mean is then given by

$$\bar{x} \pm \frac{z_{0.025} \cdot \sigma}{\sqrt{n}} = 41.924 \pm \frac{1.96 \cdot 0.1}{\sqrt{10}} = [41.862, 41.986].$$

One-Sided Confidence Intervals

13.9. Definition. Let $0 \leq \alpha \leq 1$. A $100(1 - \alpha)\%$ **upper confidence bound** for θ is a number L such that

$$P[\theta \leq L] = 1 - \alpha.$$

A $100(1 - \alpha)\%$ **lower confidence bound** for θ is a number L such that

$$P[L \leq \theta] = 1 - \alpha.$$

The corresponding intervals are called **one-sided confidence intervals**.

13.10. Theorem. Let X_1, \dots, X_n be a random sample of size n from a normal distribution with mean μ and variance σ^2 .

- (i) A $100(1 - \alpha)\%$ upper confidence bound on μ is given by $\bar{X} + \frac{z_{\alpha} \cdot \sigma}{\sqrt{n}}$.
- (ii) A $100(1 - \alpha)\%$ lower confidence bound on μ is given by $\bar{X} - \frac{z_{\alpha} \cdot \sigma}{\sqrt{n}}$.

Interval Estimation

Mathematica has built-in functionality for two-sided confidence intervals:

```
Needs["HypothesisTesting`"]

data := {41.60, 41.48, 42.34, 41.95, 41.86,
        42.18, 41.72, 42.26, 41.81, 42.04}

Mean[data]

41.924

MeanCI[data, KnownVariance -> 0.01, ConfidenceLevel -> 0.95]

{41.862, 41.986}
```

The value for $z_{\alpha/2}$ may be found by inverting the cumulative distribution function. For instance, for $\alpha = 0.05$,

```
InverseCDF[NormalDistribution[0, 1], 0.975]

1.95996
```

This is useful for finding one-sided confidence intervals.

Joint Sampling of Mean and Variance

Our interest in the chi-squared distribution is not merely abstract, for understanding the sum of squares of normally distributed random variables; in fact, the main application lies in analyzing the distribution of the sample variance. In the previous chapter, we were able to analyze the sample mean, and also its distribution, under the assumption of **known variance**. If the variance

$$\sigma^2 = E[(X - \mu)^2]$$

is unknown, we must start all over again, and first learn more about the sample variance

$$S^2 = \frac{1}{n-1} \sum_{k=1}^n (X_k - \bar{X})^2.$$

The problem essentially is that we are using the random sample X_1, \dots, X_n to obtain \bar{X} and S^2 at the same time, i.e., we actually need to obtain the **joint distribution** of \bar{X} and S^2 .

Joint Distribution of Sample Mean and Variance

The following theorem and the chi-squared distribution were discovered by Helmert in 1876 in the context of statistics of geodesical measurements. It was published in German textbooks.

However, his results were unknown to English statisticians and the chi-squared distribution was rediscovered by Pearson in 1900. Fisher and Gosset (see below) then found its application to statistics.

13.11. Theorem. Let X_1, \dots, X_n , $n \geq 2$, be a random sample of size n from a normal distribution with mean μ and variance σ^2 . Then

- (i) The sample mean \bar{X} is independent of the sample variance S^2 ,
- (ii) \bar{X} is normally distributed with mean μ and variance σ^2/n ,
- (iii) $(n-1)S^2/\sigma^2$ is chi-squared distributed with $n-1$ degrees of freedom.



Friedrich Robert Helmert (1843-1917)
File:F-R Helmert 1.jpg. (2016, January 27).
Wikimedia Commons, the free media repository.

The Helmert Transformation

The **Helmert transformation** is a very special kind of **linear, orthogonal map** from a set of $n \geq 2$ i.i.d. normal random variables X_1, \dots, X_n to a new set of random variables Y_1, \dots, Y_n .

A sample of size n taken from a normal population X with mean μ and variance σ^2 is transformed as follows:

$$\begin{aligned} Y_1 &= \frac{1}{\sqrt{n}}(X_1 + \cdots + X_n) \\ Y_2 &= \frac{1}{\sqrt{2}}(X_1 - X_2) \\ Y_3 &= \frac{1}{\sqrt{6}}(X_1 + X_2 - 2X_3) \\ &\vdots \\ Y_n &= \frac{1}{\sqrt{n(n-1)}}(X_1 + X_2 + \cdots + X_{n-1} - (n-1)X_n) \end{aligned} \tag{13.3}$$

The Helmert Transformation

In matrix notation,

$$\begin{pmatrix} Y_1 \\ Y_2 \\ Y_3 \\ \vdots \\ Y_n \end{pmatrix} = \begin{pmatrix} \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \frac{1}{\sqrt{n}} & \cdots & \frac{1}{\sqrt{n}} \\ \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} & 0 & \cdots & 0 \\ \frac{1}{\sqrt{6}} & \frac{1}{\sqrt{6}} & -\frac{2}{\sqrt{6}} & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \frac{1}{\sqrt{n(n-1)}} & \cdots & -\frac{n-1}{\sqrt{n(n-1)}} \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \\ X_3 \\ \vdots \\ X_n \end{pmatrix}$$

or $\mathbf{Y} = A\mathbf{X}$ for short. It is easy to see that the rows of the matrix A are orthonormal. Thus, A is an orthogonal matrix, $A^{-1} = A^T$. This immediately implies $|\det A| = 1$, since

$$\det A = \det A^T = \det A^{-1} = \frac{1}{\det A} \Rightarrow (\det A)^2 = 1$$

The Helmert Transformation

Incidentally, the orthogonality of A also implies that if $\mathbf{y} = A\mathbf{x}$, then

$$\sum_{i=1}^n y_i^2 = \langle \mathbf{y}, \mathbf{y} \rangle = \langle A\mathbf{x}, A\mathbf{x} \rangle = \langle A^T A\mathbf{x}, \mathbf{x} \rangle = \langle \mathbf{x}, \mathbf{x} \rangle = \sum_{i=1}^n x_i^2. \quad (13.4)$$

We have assumed that the random variables X_1, \dots, X_n are i.i.d., so their joint distribution function is given by the product of the individual normal distributions,

$$\begin{aligned} f_{X_1 \dots X_n}(x_1, \dots, x_n) &= \prod_{i=1}^n (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(x_i - \mu)^2} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \sum_{i=1}^n (x_i^2 - 2\mu x_i + \mu^2)} \\ &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n x_i^2 - 2\mu \sum_{i=1}^n x_i + n\mu^2 \right)} \end{aligned}$$

The Helmert Transformation

The Helmert transformation is linear, so its derivative (Jacobian) DA is simply A . Using (13.4), $|\det A^{-1}| = 1$ and Theorem 10.1 on the transformation of joint random variables, we obtain

$$\begin{aligned}
 & f_{Y_1 \dots Y_n}(y_1, \dots, y_n) \\
 &= f_{Y_1 \dots Y_n}(\mathbf{y}) = f_{X_1 \dots X_n}(\mathbf{x})_{\mathbf{x}=A^{-1}\mathbf{y}} \cdot \underbrace{|\det DA^{-1}(\mathbf{y})|}_{=1} \\
 &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \left(\sum_{i=1}^n y_i^2 - 2\mu\sqrt{n}y_1 + n\mu^2 \right)} \\
 &= (2\pi)^{-n/2} \sigma^{-n} e^{-\frac{1}{2\sigma^2} \left(\sum_{i=2}^n y_i^2 + (y_1 - \sqrt{n}\mu)^2 \right)} \\
 &= (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2} (y_1 - \sqrt{n}\mu)^2} \prod_{i=2}^n (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2} y_i^2}
 \end{aligned}$$

The Helmert Transformation

We see that the marginal densities are given by

$$f_{Y_1}(y_1) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}(y_1 - \sqrt{n}\mu)^2},$$

$$f_{Y_i}(y_i) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{1}{2\sigma^2}y_i^2}, \quad i = 2, \dots, n$$

and the joint density is the product of the marginal densities,

$$f_{Y_1 \dots Y_n}(y_1, \dots, y_n) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2) \dots f_{Y_n}(y_n).$$

In particular, the random variables Y_1, \dots, Y_n are **independent** and **normally distributed**.

The random variable Y_1 is normally distributed with mean $\sqrt{n}\mu$ and variance σ^2 , while Y_2, \dots, Y_n have mean 0 and variance σ^2 .

The Helmert Transformation

Proof of Theorem 13.11.

Using the Helmert transformation, we may write

$$\bar{X} = \frac{1}{\sqrt{n}} Y_1.$$

Furthermore,

$$\begin{aligned}(n-1)S^2 &= \sum_{i=1}^n (X_i - \bar{X})^2 = \sum_{i=1}^n X_i^2 - n\bar{X}^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 \\ &= \sum_{i=2}^n Y_i^2.\end{aligned}$$

Since the Y_i are all independent, it follows that \bar{X} is independent of S^2 , so we have proven the first assertion of the theorem.

The Helmert Transformation

Proof of Theorem 13.11 (continued).

Since $\bar{X} = \frac{1}{\sqrt{n}} Y_1$ and $f_{Y_1}(y_1) = (2\pi)^{-1/2} \sigma^{-1} e^{-\frac{(y_1 - \mu)^2}{2\sigma^2}}$, it follows from Theorem 7.5 that

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

so \bar{X} is normally distributed with mean μ and variance σ^2/n .

Now

$$(n-1)S^2/\sigma^2 = \frac{1}{\sigma^2} \sum_{i=2}^n Y_i^2 = \sum_{i=2}^n \left(\frac{Y_i}{\sigma}\right)^2$$

is the sum of $n-1$ squares of standard normal distributions Y_i/σ , so it follows a chi-squared distribution with $n-1$ degrees of freedom.

This completes the proof. □

Independence of Sample Mean and Sample Variance

13.12. Remark. Theorem 13.11 essentially uses the fact that the i.i.d. variables X_k , $k = 1, \dots, n$, are normally distributed. In fact, the converse result is true also:

Let X_1, \dots, X_n , $n \geq 2$, be i.i.d. random variables. Then if \bar{X} and S^2 are independent, the X_k , $k = 1, \dots, n$ follow a normal distribution.

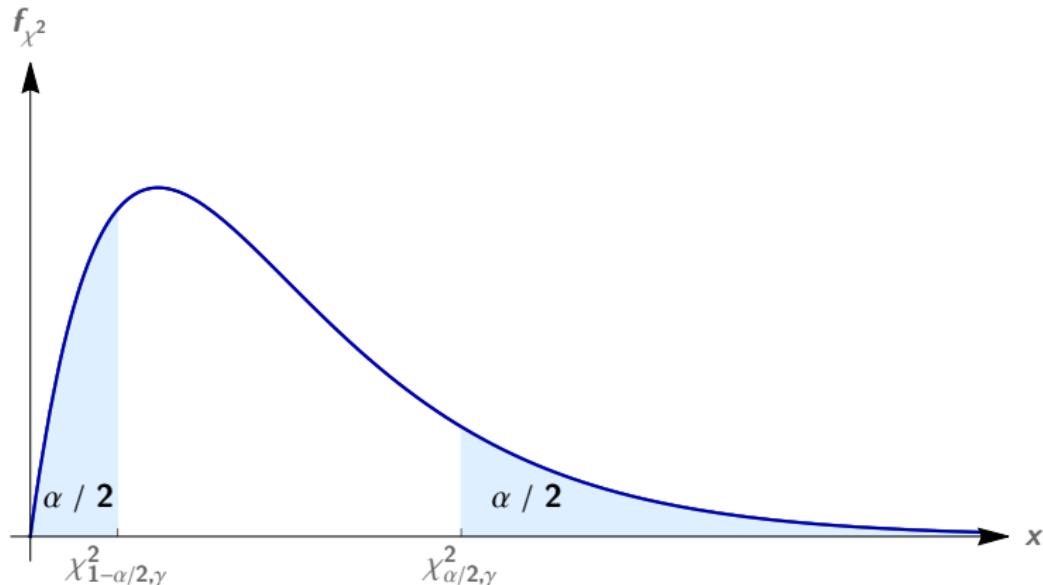
This means that the independence of \bar{X} and S^2 is a **characteristic property** of the normal distribution. Furthermore, if in a given situation we assume that \bar{X} and S^2 are independently distributed we are essentially assuming that the population is normally distributed.

We can use Theorem 13.11 to find a confidence interval for the variance based on the sample variance S^2 .

The Points $\chi^2_{1-\alpha/2,\gamma}$ and $\chi^2_{\alpha/2,\gamma}$

Given $\alpha \in [0, 1]$ and $\gamma > 0$ we define $\chi^2_{1-\alpha/2,\gamma}, \chi^2_{\alpha/2,\gamma} \in [0, \infty)$ by

$$\int_0^{\chi^2_{1-\alpha/2,\gamma}} f_{\chi^2_\gamma}(x) dx = \int_{\chi^2_{\alpha/2,\gamma}}^\infty f_{\chi^2_\gamma}(x) dx = \alpha/2,$$



Interval Estimation of Variability

From Theorem 13.11 we know that given a sample of size n from a normal population, $(n - 1)S^2/\sigma^2$ follows a chi-squared distribution with $n - 1$ degrees of freedom. Thus

$$\begin{aligned}1 - \alpha &= P\left[\chi_{1-\alpha/2, n-1}^2 \leq \frac{(n-1)S^2}{\sigma^2} \leq \chi_{\alpha/2, n-1}^2\right] \\&= P\left[\frac{(n-1)S^2}{\chi_{\alpha/2, n-1}^2} \leq \sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha/2, n-1}^2}\right]\end{aligned}$$

This gives us the following result:

13.13. Theorem. Let X_1, \dots, X_n , $n \geq 2$, be a random sample of size n from a normal distribution with mean μ and variance σ^2 . A $100(1 - \alpha)\%$ confidence interval on σ^2 is given by

$$\left[(n-1)S^2/\chi_{\alpha/2, n-1}^2, (n-1)S^2/\chi_{1-\alpha/2, n-1}^2\right].$$

Interval Estimation of Variability

Often, we are only interested in finding an upper or lower bound for the variance.

13.14. Theorem. Let X_1, \dots, X_n , $n \geq 2$, be a random sample of size n from a normal distribution with mean μ and variance σ^2 . Then with $100(1 - \alpha)\%$ confidence,

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi_{1-\alpha, n-1}^2}$$

and $[0, \frac{(n-1)S^2}{\chi_{1-\alpha, n-1}^2}]$ is a $100(1 - \alpha)\%$ **upper confidence interval** for σ^2 .

Similarly, with $100(1 - \alpha)\%$ confidence

$$\frac{(n-1)S^2}{\chi_{\alpha, n-1}^2} \leq \sigma^2.$$

and $[\frac{(n-1)S^2}{\chi_{\alpha, n-1}^2}, \infty)$ is a $100(1 - \alpha)\%$ **lower confidence interval** for σ^2 .

Interval Estimation of Variability

13.15. Example. A manufacturer of soft drink beverages is interested in the uniformity of the machine used to fill cans. Specifically, it is desirable that the standard deviation σ of the filling process be less than 0.2 fluid ounces; otherwise there will be a higher than allowable percentage of cans that are underfilled. We will assume that fill volume is approximately normally distributed. A random sample of 20 cans results in a sample variance of $s^2 = 0.0225 \text{ (fluid ounces)}^2$. A 95% upper-confidence interval is given by

$$\sigma^2 \leq \frac{(n-1)S^2}{\chi_{0.95,n-1}^2} = \frac{19 \cdot 0.0225 \text{ (fluid ounces)}^2}{10.117} = 0.0423 \text{ (fluid ounces)}^2$$

This corresponds to $\sigma \leq 0.21$ fluid ounces with 95% confidence. This is not sufficient to support the hypothesis that $\sigma \leq 0.20$ fluid ounces so further investigation is necessary.

Interval Estimation of Variability

Mathematica has built-in functionality for two-sided confidence intervals for the variance:

```
data := {41.60, 41.48, 42.34, 41.95, 41.86,  
        42.18, 41.72, 42.26, 41.81, 42.04}  
  
VarianceCI[data, ConfidenceLevel -> .95]  
  
{0.0381879, 0.269013}
```

However, one-sided intervals need to be calculated by hand, using, for example,

```
InverseCDF[ChiSquareDistribution[19], 0.05]
```

10.117

Interval Estimation for the Mean (Variance unknown)

Recall that we have derived a formula for the confidence interval of the mean of a normal distribution using the random variable

$$Z = \frac{\bar{X} - \mu}{\sigma / \sqrt{n}}$$

which was found to be normally distributed. The Central Limit Theorem allowed us to extend this result (approximately) even to non-normal distributions, but one central difficulty remained: σ must be known!

Our main goal is to derive a general formula for a confidence interval on the mean when the value of σ is not known and must be estimated.

The difficulty lies in the fact that the distribution of

$$\frac{\bar{X} - \mu}{S / \sqrt{n}}$$

is not known.

The Student T -distribution

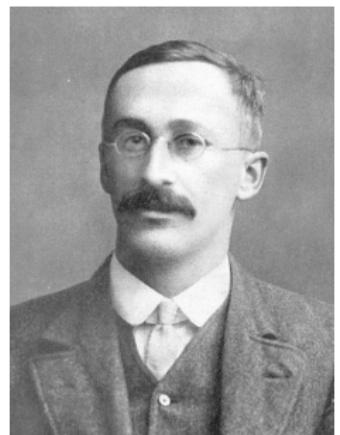
Gosset was a statistician and brewer who worked for the Guinness company in Dublin. He was interested in developing new varieties of barley and became a pioneer of many important statistical methods.

To prevent the leaking of trade secrets, Guinness prohibited their staff from publishing findings that mentioned beer, "Guinness" or the author's name. Therefore, Gosset published his results under the pseudonym "A. Student."

13.16. Definition. Let Z be a standard normal variable and let χ^2_γ be an **independent** chi-squared random variable with γ degrees of freedom. The random variable

$$T_\gamma = \frac{Z}{\sqrt{\chi^2_\gamma / \gamma}}$$

is said to follow a T -distribution with γ degrees of freedom.



William Sealy Gosset (1876-1937) In 1908.
File:William Sealy Gosset.jpg. (2017, April 26).
Wikimedia Commons, the free media repository.

Density of the T -distribution

13.17. Theorem. The density of a T distribution with γ degrees of freedom is given by

$$f_{T_\gamma}(t) = \frac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)\sqrt{\pi\gamma}} \left(1 + \frac{t^2}{\gamma}\right)^{-\frac{\gamma+1}{2}}.$$

Proof.

The distribution of χ_γ was found in (10.1) to be

$$f_{\chi_\gamma}(y) = \begin{cases} \frac{2}{2^{\gamma/2}\Gamma(\gamma/2)} y^{\gamma-1} e^{-\gamma^2/2} & y \geq 0, \\ 0 & y < 0. \end{cases}$$

It follows from Theorem 7.5 that $\sqrt{\chi_\gamma^2/\gamma} = \chi_\gamma/\sqrt{\gamma}$ has distribution

$$f_{\chi_\gamma/\sqrt{\gamma}}(y) = \begin{cases} \frac{2\sqrt{\gamma}}{2^{\gamma/2}\Gamma(\gamma/2)} (\sqrt{\gamma}y)^{\gamma-1} e^{-\gamma y^2/2} & y \geq 0, \\ 0 & y < 0. \end{cases}$$

Density of the T -distribution

Proof (continued).

The density of the standard normal random variable Z is given by

$$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}.$$

By Theorem 10.2, the density f_T of the quotient $T = X/Y$ of two independent random variables X and Y is given by

$$f_T(t) = \int_{-\infty}^{\infty} f_X(ty)f_Y(y) \cdot |y| dy.$$

For $T_\gamma = Z/(\chi_\gamma/\sqrt{\gamma})$ it follows that

$$f_{T_\gamma}(t) = \frac{1}{\sqrt{2\pi}} \frac{2\sqrt{\gamma}}{2^{\gamma/2} \Gamma(\gamma/2)} \int_0^{\infty} e^{-(t^2 + \gamma)y^2/2} (\sqrt{\gamma}y)^{\gamma-1} y dy.$$

Density of the T -distribution

Proof (continued).

Substituting $y = \sqrt{2z/(t^2 + \gamma)}$, $z = (t^2 + \gamma)y^2/2$, $dz = (t^2 + \gamma)y\,dy$, we obtain

$$\begin{aligned}f_{T_\gamma}(t) &= \frac{1}{\sqrt{2\pi}} \frac{2\sqrt{\gamma}}{2^{\gamma/2}\Gamma(\gamma/2)} (t^2 + \gamma)^{-1} \int_0^\infty e^{-z} \left(\frac{2z\gamma}{t^2 + \gamma} \right)^{\frac{\gamma-1}{2}} dz \\&= \frac{1}{\sqrt{\gamma\pi}} \frac{1}{\Gamma(\gamma/2)} \left(\frac{\gamma}{t^2 + \gamma} \right)^{\frac{\gamma+1}{2}} \int_0^\infty e^{-z} z^{\frac{\gamma+1}{2}-1} dz \\&= \frac{1}{\sqrt{\gamma\pi}} \frac{\Gamma((\gamma+1)/2)}{\Gamma(\gamma/2)} \left(1 + \frac{t^2}{\gamma} \right)^{-\frac{\gamma+1}{2}}.\end{aligned}$$

□

T-distribution of the Sample Mean

13.18. Theorem. Let X_1, \dots, X_n be a random sample from a normal distribution with mean μ and variance σ^2 . The random variable

$$T_{n-1} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

follows a *T* distribution with $n - 1$ degrees of freedom.

Proof.

We know that $(\bar{X} - \mu)/(\sigma/\sqrt{n})$ is standard normal and $(n - 1)S^2/\sigma^2$ is a chi-squared random variable with $n - 1$ degrees of freedom. Therefore,

$$\frac{Z}{\sqrt{\chi^2_\gamma/\gamma}} = \frac{(\bar{X} - \mu)/(\sigma/\sqrt{n})}{\sqrt{((n - 1)S^2/\sigma^2)/(n - 1)}} = \frac{\bar{X} - \mu}{S/\sqrt{n}}$$

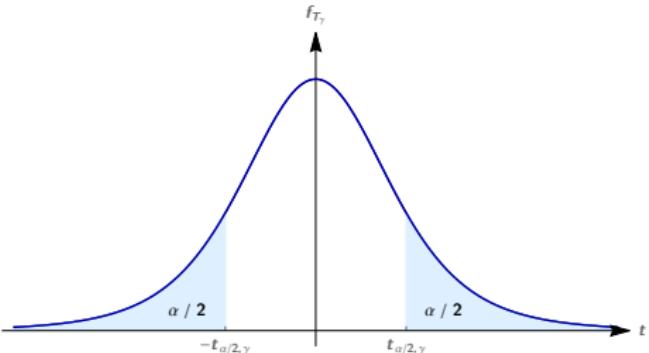
follows a *T* distribution with $n - 1$ degrees of freedom. □

Confidence Interval for the Mean (Variance Unknown)

Let $0 < \alpha \leq 1$ and $\gamma > 0$. We define $t_{\alpha/2, \gamma} \geq 0$ by

$$\int_{t_{\alpha/2, \gamma}}^{\infty} f_{T_\gamma}(t) dt = \alpha/2, \quad (13.5)$$

where f_{T_γ} is the density of the T -distribution with n degrees of freedom.



13.19. Theorem. Let X_1, \dots, X_n be a random sample of size n **from a normal distribution** with mean μ and variance σ^2 . Then a $100(1 - \alpha)\%$ confidence interval on μ is given by

$$\bar{X} \pm t_{\alpha/2, n-1} S / \sqrt{n}$$

Confidence Interval for the Mean (Variance Unknown)

13.20. Example. An article in the *Journal of Testing and Evaluation* presents the following 20 measurements on residual flame time (in seconds) of treated specimens of children's nightwear:

9.85	9.93	9.75	9.77	9.67	9.87	9.67	9.94	9.85	9.75
9.83	9.92	9.74	9.99	9.88	9.95	9.95	9.93	9.92	9.89

We wish to find a 95% confidence interval on the mean residual flame time. The sample mean and standard deviation are

$$\bar{x} = 9.8475,$$

$$s = 0.0954$$

We refer to the table for the T distribution with $20 - 1 = 19$ degrees of freedom and $\alpha/2 = 0.025$ to obtain $t_{0.025, 19} = 2.093$. Hence

$$\mu = (9.8475 \pm 0.0446) \text{ sec}, \quad \text{i.e.,} \quad 9.8029 \leq \mu \leq 9.8921$$

with 95% probability.

The Fisher Test

Hypotheses and Testing

In this section we will discuss the second major statistical method for gaining information on a probability distribution: ***hypothesis testing***. The goal is to reject or fail to reject statements (hypotheses) based on statistical data.

We will present three approaches:

- (i) Fisher's null hypothesis testing,
- (ii) Neyman–Pearson decision theory,
- (iii) The amalgam of (i) and (ii) that is still used and sometimes taught today, called Null Hypothesis Significance Testing.

In our initial discussion, a hypothesis will be a statement about a population parameter, denoted θ .

The hypothesis will compare θ to a ***null value***, denoted θ_0 .

Fisher's Null Hypothesis Test

We consider a single hypothesis that compares a population parameter θ to a given null value θ_0 .

This hypothesis will be denoted by H_0 and is called the **null hypothesis**.

Our goal is to find statistical evidence that allows us to reject the null hypothesis.

The process of using statistical data to decide whether or not a hypothesis should be rejected is called “performing a hypothesis test”.

Null hypotheses take one of three forms:

- ▶ $H_0: \theta = \theta_0$
- ▶ $H_0: \theta \leq \theta_0$
- ▶ $H_0: \theta \geq \theta_0$

Fisher's Null Hypothesis Test

14.1. Example. We want to find evidence that a new car design has a mean mileage greater than 26 mpg. Therefore, we set up the null hypothesis

$$H_0: \mu \leq 26. \quad (14.1)$$

Our goal is to gather data that allows us to **reject H_0** .

14.2. Remark. A hypothesis test is based on rejecting a hypothesis because it is possible to gather statistical evidence that a certain claim is likely to be false, while it is impossible for statistical evidence to directly prove that a claim is true. This will become more clear soon.

Suppose that the hypothesis (14.1) is given. We then take a random sample and calculate \bar{X} . If the value of \bar{X} is much greater than 26, there is reason to believe that H_0 is false.

The *P*-Value for a One-Tailed Test

The test of a hypothesis of the form

$$H_0: \theta \leq \theta_0$$

or

$$H_0: \theta \geq \theta_0$$

is said to be a **one-tailed test**.

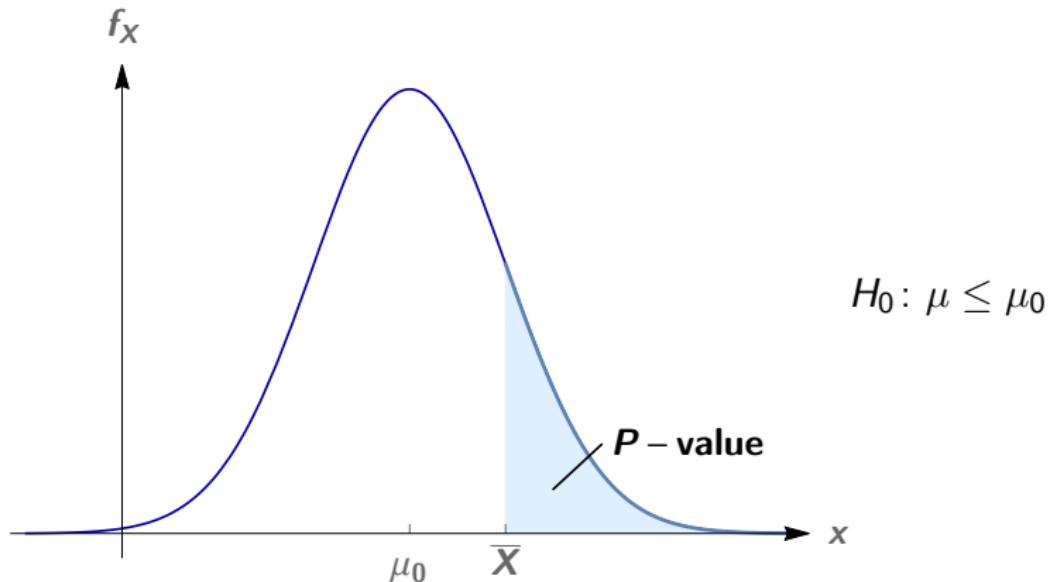
In our current example (14.1) we take a random sample of size n and find the value \bar{x} for the sample mean. We then find the probability of obtaining the measured value of \bar{x} or a larger result if $\theta = \theta_0$. This is said to be the **significance** or ***P*-value** of the test.

Note that finding the probability that we obtain \bar{x} or a greater result if $\mu = 26$ is an upper bound on the probability given $\mu \leq 26$:

$$P[\bar{X} \geq \bar{x} \mid \mu \leq 26] \leq P[\bar{X} \geq \bar{x} \mid \mu = 26]$$

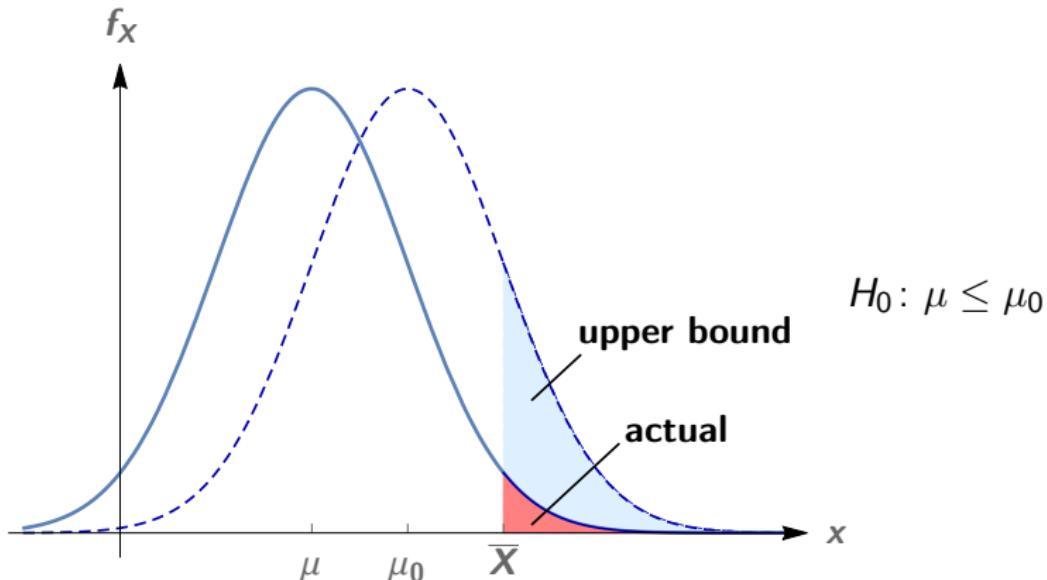
The *P*-Value for a One-Tailed Test

This is illustrated in the sketch below. The sample mean \bar{X} follows a normal distribution with mean μ . If $\mu = \mu_0$, the probability of obtaining a value of \bar{X} at least equal to the measured \bar{x} is indicated by the shaded region.



The P -Value for a One-Tailed Test

If $\mu < \mu_0$, the probability will be smaller, since the density curve will be shifted to the left.



The P -Value and Rejecting the Null Hypothesis

The P -value is therefore an upper bound of the probability of obtaining the data if H_0 is true. If D represents the statistical data,

$$P[D | H_0] \leq P\text{-value}$$

and we will reject H_0 if this value is small.

We then say that we either

- ▶ ***fail to reject H_0*** or
- ▶ ***reject H_0 at the [P-value] level of significance.***

The P -value is also called the level of significance of the test.

The statistic on which the P -value is based is called the ***test statistic***. In our discussion so far, the test statistic has been the sample mean.

A One-Tailed Test Based on the Normal Distribution

14.3. Example. Continuing from Example 14.1, we may assume that the mileage of cars currently has a standard deviation of 5 miles per gallon and that this will also be true for the new design. Furthermore, we suppose that the gas mileage follows a normal distribution.

We take a sample of 36 cars and find their gas mileages. We decide to base our rejection of H_0 on the sample mean.

If $\mu = 26$ and $\sigma = 5$, the sample mean is normally distributed with $\mu = 26$ and standard deviation $\sigma/\sqrt{n} = 5/6$.

Suppose that we find a sample mean $\bar{x} = 28.04$ mpg.

A One-Tailed Test Based on the Normal Distribution

We now calculate the P -value of the test, i.e., the probability of obtaining this or a larger value of the sample mean if H_0 were true.

$$\begin{aligned} P[\bar{X} \geq 28.04 \mid \mu \leq 26, \sigma = 5] &\leq P[\bar{X} \geq 28.04 \mid \mu = 26, \sigma = 5] \\ &= P\left[\frac{\bar{X} - 26}{5/6} \geq \frac{28.04 - 26}{5/6}\right] \\ &= P[Z \geq 2.45] = 1 - P[Z \leq 2.45] \\ &= 1 - 0.9929 = 0.0071. \end{aligned}$$

This is the P -value of the test. Since it is very small, we decide to reject the null hypothesis at the 0.7% level of significance.

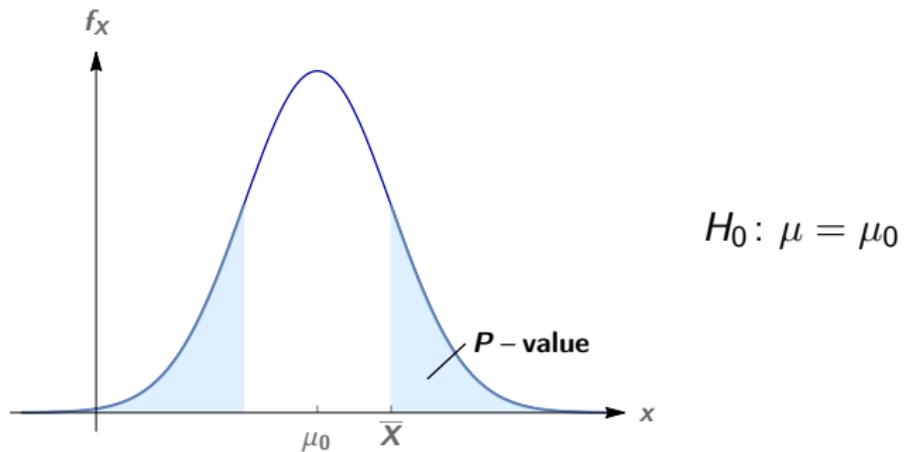
We may say that there is evidence that the gas mileage of the cars of new design is greater than 26 mpg.

Two-Tailed Tests

If we are testing a hypothesis of the form

$$H_0: \theta = \theta_0$$

we say we are performing a **two-tailed test**. In this case, the P -value is twice the value of a one-tailed test, since there is evidence that the null hypothesis is false if the statistic differs from θ_0 significantly, regardless of whether the statistic is greater or smaller.



A Two-Tailed Test Based on the Normal Distribution

14.4. Example. The burning rate of a rocket propellant is being studied. Specifications require that the mean burning rate must be 40 cm/s. Furthermore, suppose that we know that the standard deviation of the burning rate is approximately $\sigma = 2$ cm/s. The experimenter decides to base the test on a random sample of size $n = 25$. The null hypothesis is

$$H_0: \mu = 40 \text{ cm/s}$$

If H_0 is true, the sample mean is normally distributed with mean $\mu_0 = 40$ cm/s and variance σ^2/n ; thus she will use the test statistic

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

which is standard normal if H_0 is true.

The Z-Test

Twenty-five specimen are tested, and the sample mean burning rate obtained is $\bar{x} = 41.25 \text{ cm/s}$. The value of the test statistic is

$$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}} = \frac{41.25 - 40}{2/\sqrt{25}} = 3.125.$$

Then

$$P[Z \geq 3.125 | H_0] = 1 - P[Z \leq 3.125 | H_0] = 1 - 0.9991 = 0.0009$$

Since this is a two-tailed test, the P -value is twice this number and she decides to reject H_0 at the 0.18% level of significance.

There is evidence that the burning rate is not 40 cm/s.

Fisher's Null Hypothesis Test

14.5. Remarks.

- ▶ Fisher originally recommended rejecting H_0 if the P -value is less than 5%, i.e., $P < 0.05$. However, he later changed his mind and advocated quoting the actual P -value and deciding whether or not to reject H_0 on a case-by-case basis.
- ▶ According to Fisher, this type of test should only be used if very little is known about the parameter θ . The hypothesis test is just a first step in investigating θ . Confidence intervals and other techniques give far more information in practice.
- ▶ Fisher also observed that a single significant test should not be enough to comprehensively reject H_0 . Only multiple, independent significant test should be enough to allow the conclusion that H_0 is actually false.

Does a small P -value provide evidence that H_0 is false?

But there is also a more fundamental issue: what a researcher wants is, given data D , the probability that H_0 is true, i.e.,

$$P[H_0 | D].$$

But the P -value is just the converse probability, i.e.,

$$P[D | H_0].$$

It is easy to write down Bayes's theorem and see that

$$P[H_0 | D] = \frac{P[D | H_0] \cdot P[H_0]}{P[D | H_0] \cdot P[H_0] + P[D | \neg H_0] \cdot P[\neg H_0]}.$$

Since $P[\neg H_0] = 1 - P[H_0]$ we find that

$$P[H_0 | D] = \frac{P[D | H_0] \cdot P[H_0]}{P[D | H_0] \cdot P[H_0] + P[D | \neg H_0](1 - P[H_0])}.$$

Is Hypothesis Testing logical?

Then if $P[H_0] \neq 0$ then

$$\begin{aligned} P[H_0 | D] &= \frac{P[D | H_0] \cdot P[H_0]}{P[D | H_0] \cdot P[H_0] + P[D | \neg H_0](1 - P[H_0])} \\ &= \frac{1}{1 + \frac{P[D | \neg H_0]}{P[D | H_0]} \frac{1 - P[H_0]}{P[H_0]}} \end{aligned}$$

This shows the following:

- ▶ If $P[H_0]$ is small, even a large P -value does not mean that H_0 is likely to be true.
- ▶ If $P[H_0]$ is large, even a small P -value does not mean that H_0 is likely to be false.

Hence, it is possible that we have data which is very unlikely given H_0 , but that in fact H_0 given the data is very likely (and vice-versa).

In short, classical hypothesis testing does not take the probability of H_0 being true in the first place into account.

Bayesian vs. Frequentist Statistics

This problem has given rise to ***Bayesian statistics*** which attempts to assign a ***prior probability*** to H_0 and deduce a ***posterior probability*** based on experimental results. However, finding this prior probability is often tricky. There are, broadly speaking, two groups of statisticians:

- ▶ **Frequentists**, who mainly ignore the problems mentioned here or claim that they are not relevant in their specific research (for example, because they consider $P[H_0]$ to not be small in their experiments).
- ▶ **Bayesians** who claim to understand the logical inconsistencies and intend to compensate for them via prior and posterior probability distributions. While theoretically pure, this may be difficult to implement in practice.

Of course, these are extreme characterizations. In practice, every statistician knows Bayes's theorem and will apply it as much as possible and no statistician entirely rejects frequentist methods.

Bayesian vs. Frequentist Statistics

DID THE SUN JUST EXPLODE?
(IT'S NIGHT, SO WE'RE NOT SURE.)

THIS NEUTRINO DETECTOR MEASURES
WHETHER THE SUN HAS GONE NOVA.

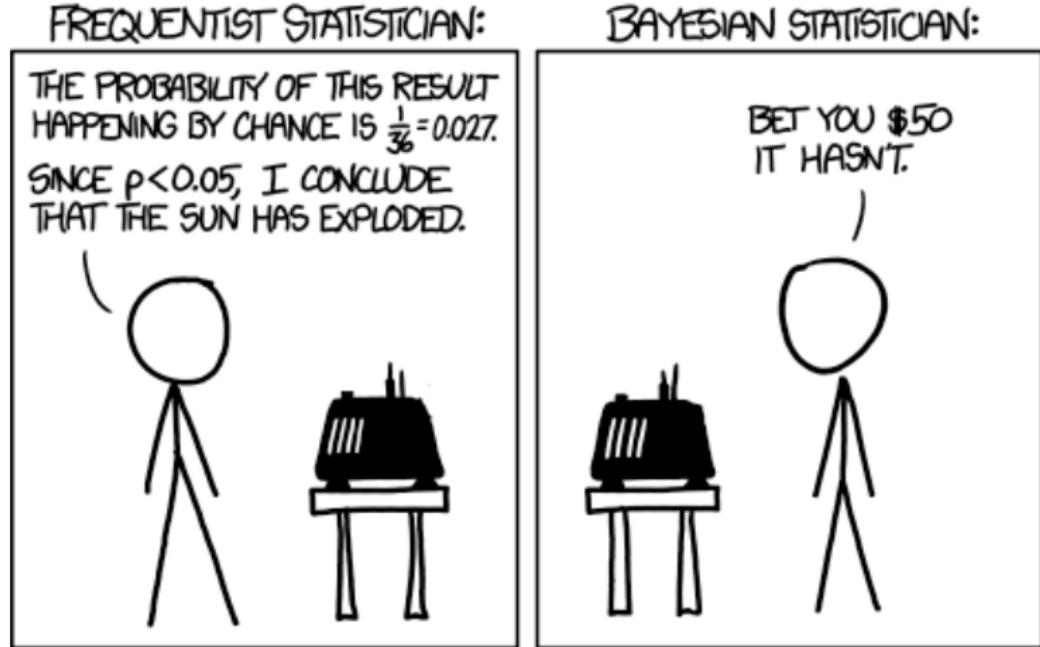
THEN, IT ROLLS TWO DICE. IF THEY
BOTH COME UP SIX, IT LIES TO US.
OTHERWISE, IT TELLS THE TRUTH.

LET'S TRY.

DETECTOR! HAS THE
SUN GONE NOVA?



Bayesian vs. Frequentist Statistics



xkcd: frequentists vs. bayesians, Randall Munroe, published on xkcd.com, September 11, 2012

Does a small P -value provide evidence that H_0 is false?

A more serious example:

For young women of age 30 the incidence of live-born infants with Down's syndrome is 1/885, and the majority of pregnancies are normal. Even if the two conditional probabilities of a correct test result, given either an affected or a normal fetus, were 99.5 percent, the probability of an affected child, given a positive test result, would be only 18 percent. [...]

Thus, if we substitute "The fetus is normal" for H_0 , and "The test result is positive (i.e. indicating Down's syndrome)" for D , we have $P[D | H_0] = 0.005$, which means D is a significant result, while $P[H_0 | D] = 0.82$ (i.e., $1 - 0.18$).

Pauker, S. P., & Pauker, S. G. (1979). *The amniocentesis decision: An explicit guide for parents*. In C. J. Epstein, C. J. R. Curry, S. Packman, S. Sherman, & B. D. Hall (Eds.), *Birth defects: Original article series: Volume 15. Risk, communication, and decision making in genetic counseling* (pp. 289-324). New York: The National Foundation.

Is Rejecting H_0 Trivial?

Tukey and others have argued that a null hypothesis of the form $H_0: \mu = \mu_0$ is never true in practice; at some point in the decimal expansion of the null value and the (unknown) true value, a difference will occur with probability 1.

Therefore, testing to reject H_0 is pointless: a significant result can always be obtained if the sample size n is chosen large enough. Conversely, a failure to reject H_0 simply means that the sample size wasn't large enough.

Hence, if H_0 is rejected, that does not show that H_0 was false (by the above argument this was obvious anyway) but only that the researcher was clever enough to put together a test with enough power to detect this.

One solution for this problem is to avoid two-tailed tests entirely.

Interpretation of the P -value

Suppose that you perform a Fisher test comparing a mean μ to a null value μ_0

$$H_0: \mu \leq \mu_0.$$

After obtaining your data and from it the sample mean \bar{X} , you find that $\bar{X} > \mu_0$ and calculate a P -value of 0.3%.

Which of the following statements are correct?

- (1) There is at least a 99.7% chance that H_0 is true.
- (2) There is at least a 0.3% chance that H_0 is true.
- (3) If H_0 were true, there would be at most a 0.3% chance of obtaining a value of \bar{X} equal or greater to the one measured.
- (4) If H_0 were false, there would be at least a 99.7% chance of obtaining a value of \bar{X} equal to the one measured or greater.

Neyman-Pearson Decision Theory

Neyman-Pearson Decision Theory

In Neyman-Pearson decision theory, we consider two competing hypotheses, denoted H_0 and H_1 .

As before, we seek to **reject H_0** , in which case we **accept H_1** .

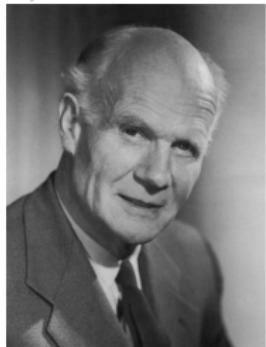
We say that

- ▶ H_0 is the **null hypothesis**,
- ▶ H_1 is the **research hypothesis** or **alternative hypothesis**.

The main difference to Fisher's approach is that we actually want to make a decision between two discrete possibilities instead of just finding evidence for or against H_0 .



Neyman, Jerzy (1894-1981) Jerzy Neyman, Book of Proofs, <https://www.bookofproofs.org/history/jerzy-neyman/>



Egon Sharpe Pearson (1895-1980)
Bartlett, M. S. Egon Sharpe Pearson.
11 August 1895-12 June 1980.
Biographical Memoirs of Fellows of the Royal Society, vol. 27, 1981, pp. 425443. JSTOR

Example of Neyman-Pearson Decision Theory

15.1. Example. Let us revisit Example 14.4. The mean burning rate for a rocket propellant is supposed to be $\mu_0 = 40 \text{ cm/s}$. It is known that the standard deviation is $\sigma = 2 \text{ cm/s}$. If the rocket propellant burns significantly too fast or too slowly, it can not be used. An experimenter sets out the two hypotheses

$$H_0: \mu = 40,$$

$$H_1: |\mu - 40| \geq 1.$$

If there is evidence that H_1 is true, the rocket propellant must be discarded, otherwise it can be used.

The P -value in Fisher's test procedure represents a continuum of evidence against H_0 , while in the Neyman-Pearson approach we will define a sharp cut-off point for our data. If the data lies beyond this cut-off point, H_0 is rejected and H_1 is accepted.

Accepting Hypotheses

The statistical test will end with either

- ▶ failing to reject H_0 , therefore accepting H_0 or
- ▶ rejecting H_0 , thereby accepting H_1 .

If we accept H_0 , we do not necessarily believe H_0 to be true; we simply decide to act as if it were true. The same is the case if we decide to accept H_1 ; we are not necessarily convinced that H_1 is true, we merely decide to assume that it is.

15.2. Example. In the situation described in Example 15.1,

- ▶ accepting H_0 means that we assume that the rocket propellant burns at a mean rate of 40 cm/s. It does not mean that we actually believe that the value is precisely 40 and not 39.993, for instance.
- ▶ accepting H_1 means that we assume that the rocket fuel burns at a rate different by more than 1 cm/s from the nominal rate. It does not necessarily mean that we have evidence to support this, merely that we will assume that it is the case.

Type I and Type II Errors

Given a choice between H_0 and H_1 , there are four possible outcomes of the decision-making process:

- (i) We reject H_0 (and accept H_1) when H_0 is false.
- (ii) We reject H_0 (accept H_1) even though H_0 is true (**Type I error**).
- (iii) We fail to reject H_0 even though H_1 is true (**Type II error**).
- (iv) We fail to reject H_0 when H_0 is true.

We will design a test to decide between rejecting or failing to reject H_0 based solely on the probability of committing Type I or Type II errors, which we want (of course) to keep as small as possible.

Power, Type I & Type II Error Probabilities

We define the probability of committing a Type I error,

$$\begin{aligned}\alpha &:= P[\text{Type I error}] = P[\text{reject } H_0 \mid H_0 \text{ true}] \\ &= P[\text{accept } H_1 \mid H_0 \text{ true}].\end{aligned}$$

The probability of committing a Type II error is denoted

$$\begin{aligned}\beta &:= P[\text{Type II error}] = P[\text{fail to reject } H_0 \mid H_1 \text{ true}] \\ &= P[\text{accept } H_0 \mid H_1 \text{ true}].\end{aligned}$$

Related to β is the **power** of the test, defined as

$$\begin{aligned}\text{Power} &:= 1 - \beta = P[\text{reject } H_0 \mid H_1 \text{ true}] \\ &= P[\text{accept } H_1 \mid H_1 \text{ true}].\end{aligned}$$

α and the Critical Region

To set up the test, we select a test statistic and determine a **critical region** for the test: if the value of the test statistic falls into the critical region, then we reject H_0 . Our critical region is determined by the desire to keep α small, e.g., less than 5%.

Hence, we determine the critical region in such a way that if H_0 is true, then the probability of the test statistic's values falling into the critical region is not more than α .

15.3. Example. In the situation described in Example 15.1, we may use \bar{X} as a test statistic. The experimenter tests a sample of $n = 25$ specimen.

If H_0 is true, \bar{X} will follow a normal distribution with mean $\mu = 40$ and $\sigma/\sqrt{n} = 2/5$, i.e.,

$$Z = \frac{\bar{X} - \mu_0}{\sigma/\sqrt{n}}$$

follows a standard normal distribution.

α and the Critical Region

Hence, with a probability of $1 - \alpha$,

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}.$$

If H_0 is true, then the probability that

$$\frac{|\bar{X} - \mu_0|}{\sigma/\sqrt{n}} > z_{\alpha/2}$$

is equal to α . Therefore, the critical region is determined by

$$\bar{x} \neq \mu_0 \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}. \quad (15.1)$$

α and the Critical Region

Suppose the experimenter would like to limit α , the probability of committing a Type I error if she rejects H_0 , to 5%. This corresponds to $z_{\alpha/2} = 1.96$ and inserting the values for μ_0 , σ and n , we find with probability $1 - \alpha$,

$$39.216 < \bar{X} < 40.784.$$

Hence the **critical region** is determined by

$$|\bar{X} - 40| > 0.784. \quad (15.2)$$

If \bar{X} falls into the range of values satisfying (15.2), the experimenter will reject H_0 , knowing that this decision will be wrong with a probability of at most 5%.

α and the Critical Region

15.4. Remarks.

- (i) In this scheme, The decision whether to reject H_0 or not is not driven by the probability of H_0 being true or not, but solely by the probability of committing an error if H_0 is falsely rejected.
- (ii) Only H_0 plays a role in the calculation of the critical region. H_1 does not enter into the discussion at all.
- (iii) Rejecting H_0 (when the data falls into the critical region) does not actually mean that there is proof that H_1 is true; in the example above, H_0 can be rejected even if $|\bar{X} - 40| < 1$.

α and the Critical Region

If the experimenter in the previous example had wanted to reduce the probability of making a wrong decision when rejecting H_0 , she could have set a higher bar for rejection: to achieve $\alpha = 1\%$, she would require

$$\left| \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}} \right| \geq z_{\alpha/2} = 2.575.$$

This would lead to a critical region of

$$|\bar{X} - 40| > 1.03.$$

If H_0 were then rejected because the sample mean fell into the critical region, the chance of this being in error would only be 1%. The trade-off is that it becomes less likely that the data will allow rejection of H_0 in the first place.

In this context, it is important to note:

*In order for the statistical procedure to be valid, the critical region must be fixed **before data are obtained**.*

β and the Sample Size

The second type of error concerns failing to reject H_0 even though H_1 is true. We calculate this probability in the case of

$$H_0: \mu = \mu_0,$$

$$H_1: |\mu - \mu_0| \geq \delta_0$$

as follows. Suppose that the true value of the mean is $\mu = \mu_0 + \delta$, $\delta > 0$.
The test statistic

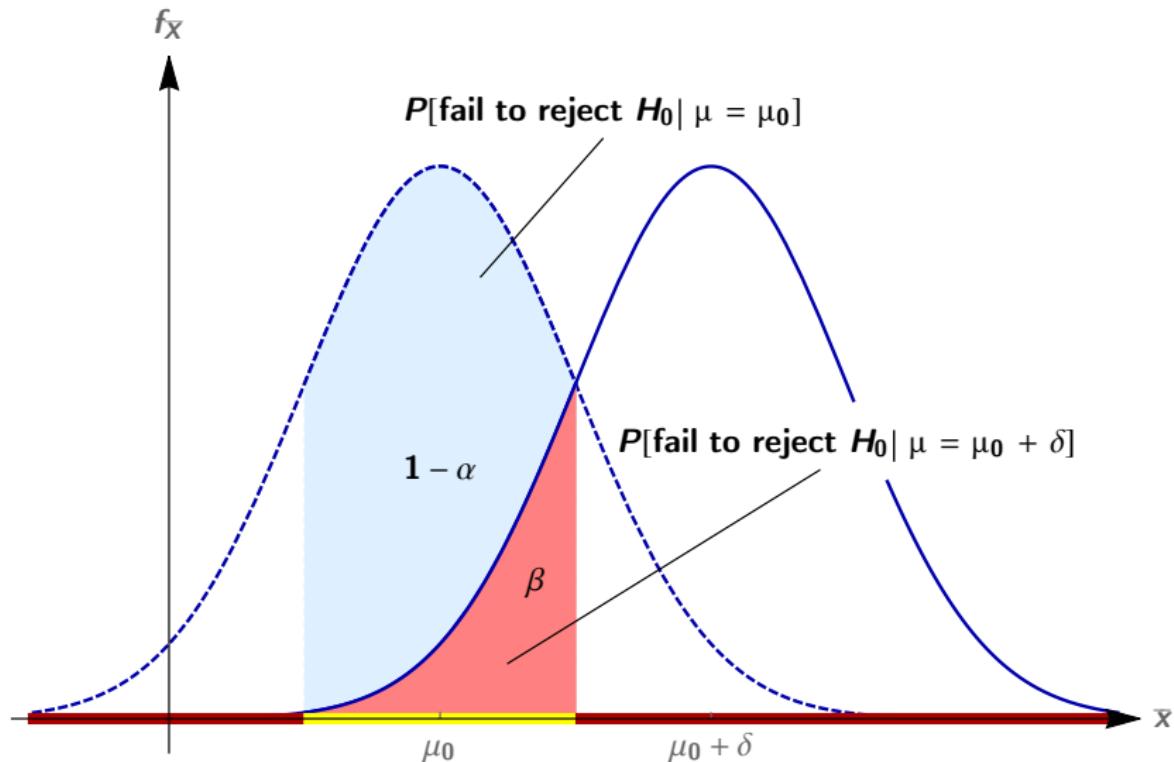
$$Z = \frac{\bar{X} - \mu_0}{\sigma / \sqrt{n}}$$

will then follow a normal distribution with unit variance and mean $\delta \sqrt{n} / \sigma$.
Supposing that α has been fixed, we will **fail to reject H_0** if

$$-z_{\alpha/2} \leq Z \leq z_{\alpha/2}.$$

or

$$\mu_0 - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{x} < \mu_0 + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

Illustration of β 

Calculating β for the Normal Distribution

Using the density of the normal distribution, we then find

$$\begin{aligned} & P[\text{fail to reject } H_0 \mid \mu = \mu_0 + \delta] \\ &= \frac{1}{\sqrt{2\pi}} \int_{-z_{\alpha/2}}^{z_{\alpha/2}} e^{-(t-\delta\sqrt{n}/\sigma)^2/2} dt \\ &= \frac{1}{\sqrt{2\pi}} \int_{-z_{\alpha/2}-\delta\sqrt{n}/\sigma}^{z_{\alpha/2}-\delta\sqrt{n}/\sigma} e^{-t^2/2} dt \quad (15.3) \\ &\approx \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2}-\delta\sqrt{n}/\sigma} e^{-t^2/2} dt. \end{aligned}$$

Calculating β for the Normal Distribution

Let us suppose H_1 is true, i.e., $|\mu - \mu_0| \geq \delta_0$. Then

$$\begin{aligned}\beta &= P[\text{fail to reject } H_0 \mid H_1 \text{ true}] \\ &\leq P[\text{fail to reject } H_0 \mid \mu = \mu_0 + \delta_0]\end{aligned}$$

and we have (to good approximation)

$$\beta(\mu) \leq \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{z_{\alpha/2} - \delta\sqrt{n}/\sigma} e^{-t^2/2} dt.$$

Adapting the notation from (13.2), we use the number $z_\beta \in \mathbb{R}$ to indicate

$$\beta = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{-z_\beta} e^{-t^2/2} dt.$$

Calculating β for the Normal Distribution

Then the relationship between δ , α , β and n with σ known is given by

$$-z_\beta \approx z_{\alpha/2} - \delta\sqrt{n}/\sigma$$

or

$$n \approx \frac{(z_{\alpha/2} + z_\beta)^2 \sigma^2}{\delta^2}. \quad (15.4)$$

In this way a desired (small) β can be attained by choosing an appropriate sample size n .

Similarly to the convention used for α , the number β when quoted for a Neyman-Pearson test usually refers to the upper bound of committing a Type II error.

Designing an Experiment for Desired α and β

15.5. Example. Revisiting Example 15.1, the experimenter would like to test the hypotheses

$$H_0: \mu = 40,$$

$$H_1: |\mu - 40| \geq 1.$$

in such a way that $\alpha = 5\%$ and $\beta = 10\%$, i.e, if H_0 is rejected, there is a 5% chance of this being in error, and if H_0 is not rejected (H_1 is accepted) there is a 10% chance of this being in error.

The critical region is set as before and the necessary sample size is calculated from (15.4) using $\beta = 0.10$, $\alpha = 0.05$, $\sigma = 2 \text{ cm/s}$ and $\delta = 1 \text{ cm/s}$. Then

$$n \approx 42,$$

so the sample size should be at least 42 to ensure $\beta \leq 0.10$.

Power

Another way to think about β is in terms of **power**, defined as $1 - \beta$ and formally given by

$$1 - \beta = P[\text{accept } H_1 \mid H_1 \text{ true}].$$

A given experiment is set up so that we either reject H_0 or we don't. Generally, we would like the probability of rejecting H_0 if the alternative hypothesis is true to be high, i.e., β to be small. Choosing a sufficiently large sample size ensures that the data gathered is powerful enough to actually reject H_0 , assuming H_1 is true.

One says that an experiment has **high power** if rejection of H_0 is likely, assuming H_1 is true. Generally speaking, a given test is more powerful than another if it requires a smaller sample size to attain the same β .

Operating Characteristic (OC) Curves

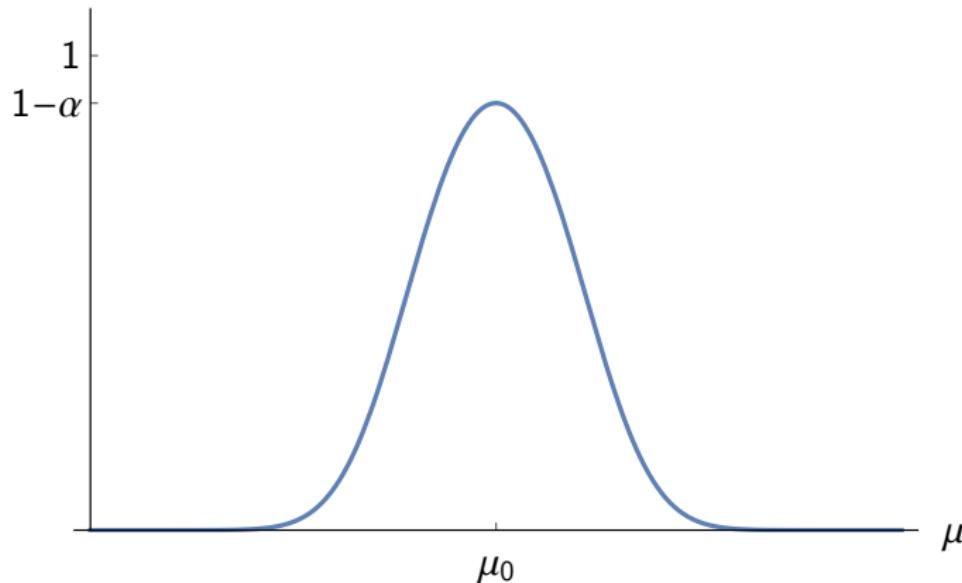
In (15.3) we calculated the probability of failing to reject H_0 as an integral. In practice, it may be difficult to perform such a calculations for non-normal distributions and evaluating the resulting integral may be impractical. For this reason, it is possible to refer to so-called ***operating characteristic curves***, known also as ***OC curves***.

A single OC curve plots the probability of failing to reject H_0 in a one-sided or two-sided test as a function of the parameter θ . A single such curve represents a choice of test parameters α and n . Other parameters of the distribution are also incorporated into the graph.

Operating Characteristic (OC) Curves

The figure below shows an OC curve for a two-sided test of the null hypothesis $H_0: \mu = \mu_0$ performed at fixed level α and fixed sample size n .

$P[\text{fail to reject } H_0]$



Effect of α on an OC Curve

Note that

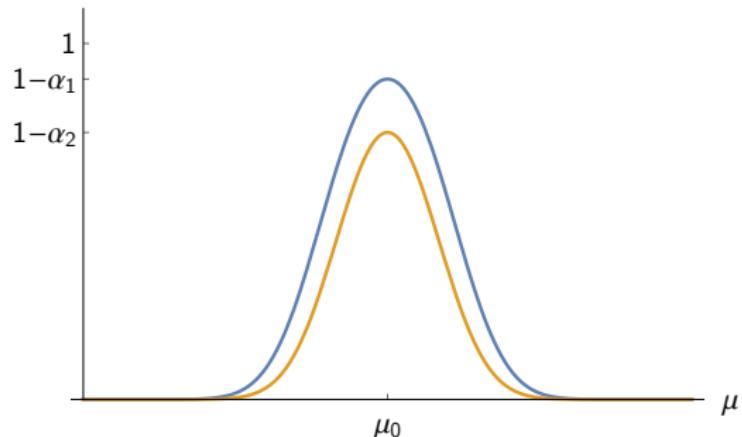
$$P[\text{fail to reject } H_0 \mid \mu = \mu_0] = 1 - \alpha,$$

since

$$P[\text{reject } H_0 \mid \mu = \mu_0] = P[\text{reject } H_0 \mid H_0 \text{ true}] = \alpha,$$

by the construction of the test. For different values of α , the curves scale correspondingly:

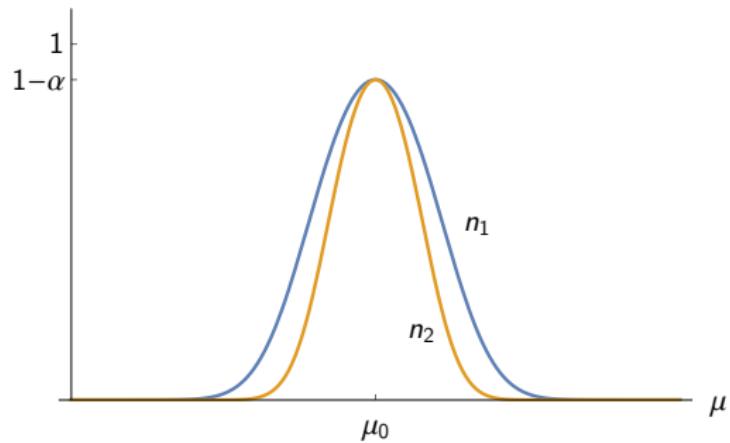
$$P[\text{fail to reject } H_0]$$



Effect of the Sample Size on an OC Curve

The sample size affects an OC curve as shown below for $n_2 > n_1$:

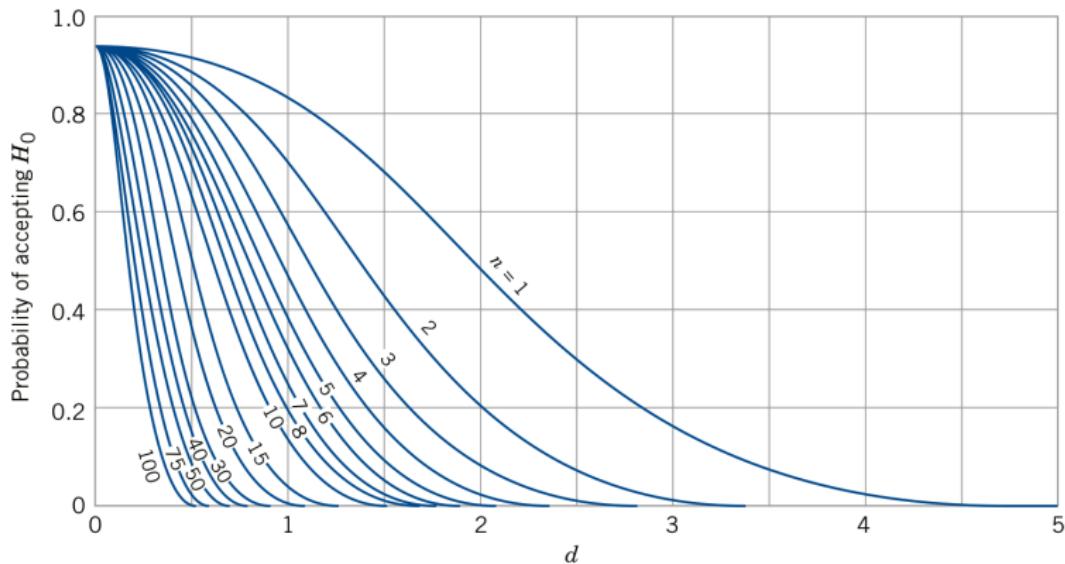
$P[\text{fail to reject } H_0]$



A typical graph will show OC curves for various values of n . Furthermore, for two-sided tests, only the right-hand half of the curve is shown to save space.

Using OC Curves to Relate Sample Sizes with β

15.6. Example. Continuing from Example 14.4, suppose that the analyst is concerned about the probability of a Type II error if the true mean burning rate is $\mu = 41 \text{ cm/s}$. We may use the following operating characteristic curve (specific to $\alpha = 0.05$) to find β :



Using OC Curves to Relate Sample Sizes with β

In this graph,

$$d := \frac{|\mu - \mu_0|}{\sigma} = \frac{41 - 40}{2} = \frac{1}{2}.$$

Since in our example $n = 25$ we can read off $\beta \approx 0.30$.

15.7. Example. In Examples 15.5 we used a formula to find the sample size necessary to reject H_0 if H_1 is actually true. We can also read the result directly from the OC curve as follows:

We want to have $\beta \leq 0.1$ if

$$d = \frac{|\mu - \mu_0|}{\sigma} = \frac{|\mu - 40|}{2} \geq \frac{1}{2}.$$

We see that the point $(d, \beta) = (0.5, 0.1)$ is between the OC curves for $n = 40$ and $n = 50$ and that the curve remains below 0.1 for $d > 1/2$. Thus the test should involve a sample size of about $n = 45$ or more.

OC Curves for One-Tailed Tests

Given a one-sided null hypothesis of the form

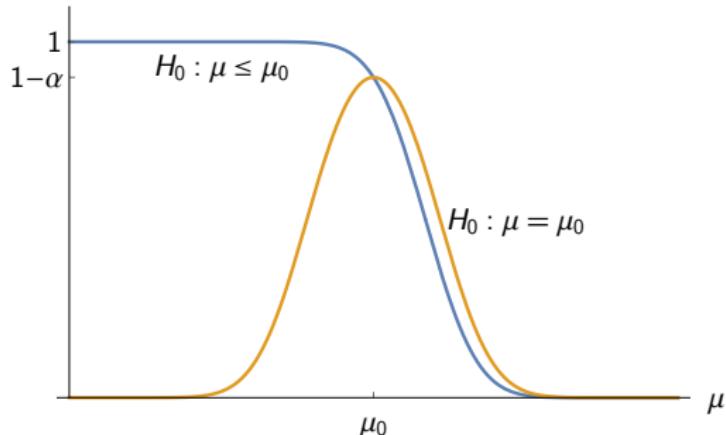
$$H_0: \theta \leq \theta_0,$$

or

$$H_0: \theta \geq \theta_0$$

an analogous calculation the probability of failing to reject H_0 may be performed, leading to an OC curve as shown below:

$P[\text{fail to reject } H_0]$



Summary of Neyman-Pearson Decision Theory

- (i) Select appropriate hypotheses H_1 and H_0 and a test statistic;
- (ii) Fix α and β for the test;
- (iii) Use α and β to determine the appropriate the sample size;
- (iv) Use α and the sample size to determine the critical region;
- (v) Obtain the sample statistic; if the test statistic falls into the critical region, reject H_0 at significance level α and accept H_1 . Otherwise, accept H_0 .

Comparison of Fisher and Neyman-Pearson Tests

Superficially, Fisher's test of H_0 and the Neyman-Pearson test are related as follows:

If the P-value in Fisher's test is no greater than the value of α in Neyman-Pearson's decision process, then H_0 is rejected and H_1 accepted. Otherwise, H_0 is not rejected.

However, this ignores the different philosophies of the approaches: Fisher is concerned about gathering evidence against H_0 , without necessarily coming to an outright rejection, while Neyman-Pearson desire a definite decision for either H_1 or H_0 .

Relationship to Confidence Intervals

We have seen in (15.1) that the two-tailed null hypothesis $H_0: \mu = \mu_0$ is rejected if

$$\bar{x} \neq \mu_0 \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

This is equivalent to

$$\mu_0 \notin \bar{x} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}}.$$

Hence, we have the following relationship to hypothesis tests:

- ▶ **Neyman-Pearson:** \bar{x} lies in the critical region for α if and only if the null value μ_0 does not lie in a $100(1 - \alpha)\%$ two-sided confidence interval for μ .
- ▶ **Fisher:** H_0 is rejected at significance level α if and only if the null value μ_0 does not lie in a $100(1 - \alpha)\%$ two-sided confidence interval for μ .

This generalizes to one-sided tests and is also true for other (non-normal) distributions.

Interpretation of the Neyman-Pearson Decision

Suppose that you are performing a Neyman-Pearson test for a population mean with

$$H_0: \mu \leq \mu_0,$$

$$H_1: \mu > \mu_1$$

where $\mu_0 < \mu_1$. The test has been designed so that $\alpha = 1\%$, $\beta = 5\%$.

Finally, H_0 is not rejected, i.e., H_0 is accepted. Then

- (1) There is at most a 5% chance that H_1 is true.
- (2) There is a 99% chance that H_0 is true.
- (3) There is a 95% chance of this conclusion being correct.
- (4) If H_1 is in fact true, the chance of reaching this conclusion is at most 5%.

Null Hypothesis Significance Testing

Null Hypothesis Significance Testing

Modern textbooks with titles such as “Statistics for Engineers” and similar do not explicitly teach either Fisher’s Test procedure nor the Neyman-Pearson decision-making process, but rather a mixture of both.

This is now often called ***Null Hypothesis Significance Testing (NHST)*** and works as follows:

- ▶ Two hypotheses, H_0 and H_1 are set up, but H_1 is always the logical negation of H_0
- ▶ Then either a “hypothesis test” is performed, whereby a critical region for given α is defined, the test statistic is evaluated and H_0 is either rejected or accepted.
- ▶ Alternatively (and more commonly), the test statistic is evaluated immediately, a P -value is found, and H_0 is either rejected or accepted based on that value.
- ▶ In either case, there is no meaningful discussion of β , since H_1 is exactly the negation of H_0 .

Criticism of NHST

- ▶ A small P -value does not guarantee that a large probability that H_0 is false. Fisher did not intend for a small P -value to lead to a clear rejection of H_0 , but only to serve as evidence against H_0 if little else is known.
- ▶ Rejecting H_0 based on $\alpha = 0.05$ or 0.01 or any other value is arbitrary.
- ▶ NHST is actually biased ***against failing to reject H_0*** . From a Bayesian point of view, it is far too easy to reject H_0 because $P[H_0]$ does not enter into NHST.
- ▶ A two-sided test such as $H_0: \theta = \theta_0$, $H_1: \theta \neq \theta_0$ is meaningless.
- ▶ The power (and β) of the test is not properly defined, since H_1 is just the alternative “not H_0 ” rather than referring to a distinct value θ_1 . Occasionally, this θ_1 is then mentioned indirectly for purposes of power calculations.

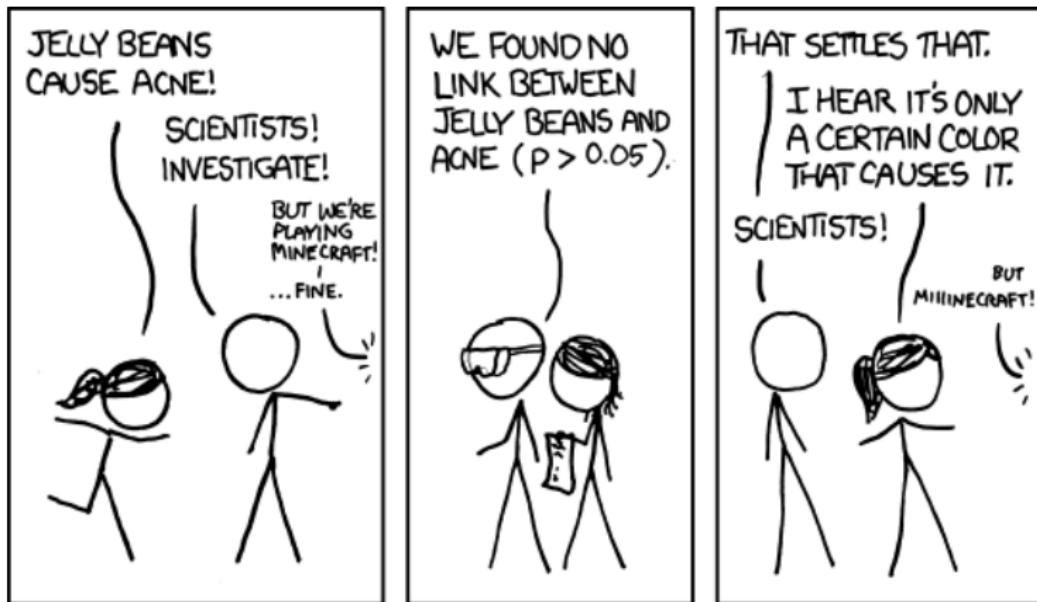
Publication Bias and NHST

NHST is currently the preferred technique for verifying statistical results. In the current academic environment, research papers are only publishable if the results are statistically significant. Editors of scientific journals will usually not publish results where $P = 0.23$, for example.

This means that many interesting studies are not made available to the scientific community because they are considered to be “failed experiments”. However, that does not mean that they are not useful (even if H_0 is true) or that H_0 actually is false (since the experiment may simply not have had enough power).

At the same time, this tempts researchers to continue increasing the sample sizes of a study or to do repeated studies until they get a result that is statistically significant and can be published.

Publication Bias and NHST



xkcd: significant, Randall Munroe, published on xkcd.com, April 6, 2011

Publication Bias and NHST

WE FOUND NO
LINK BETWEEN
PURPLE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BROWN JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
PINK JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
BLUE JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
TEAL JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
SALMON JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
RED JELLY
BEANS AND ACNE
($P > 0.05$).



WE FOUND NO
LINK BETWEEN
TURQUOISE JELLY
BEANS AND ACNE
($P > 0.05$).



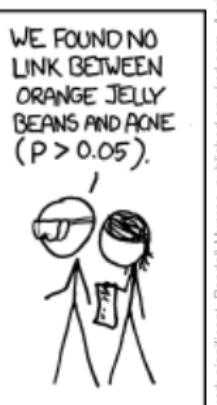
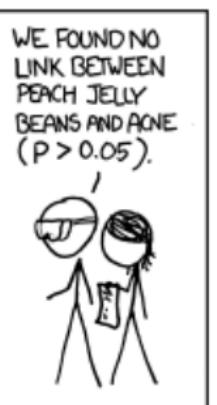
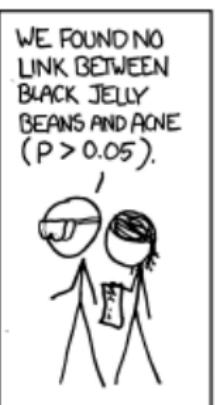
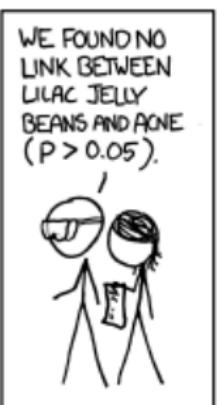
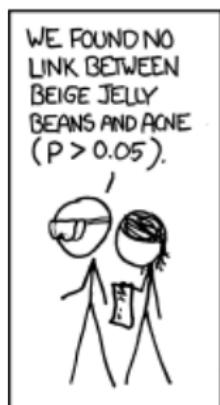
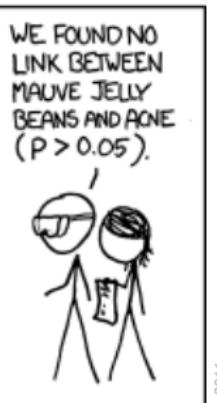
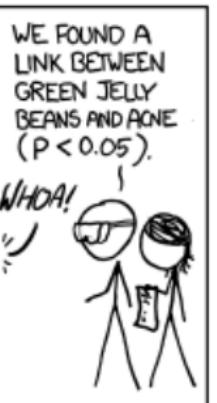
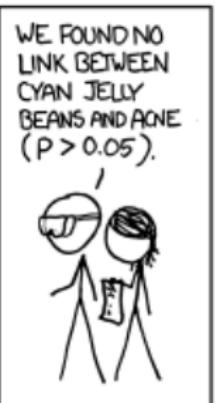
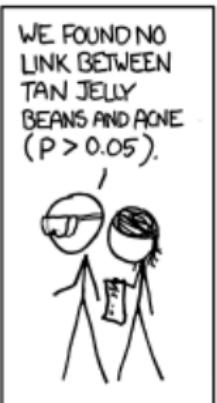
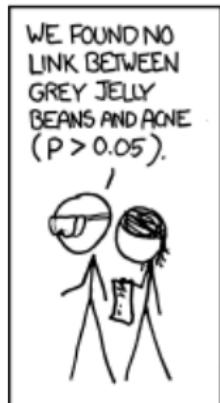
WE FOUND NO
LINK BETWEEN
MAGENTA JELLY
BEANS AND ACNE
($P > 0.05$).



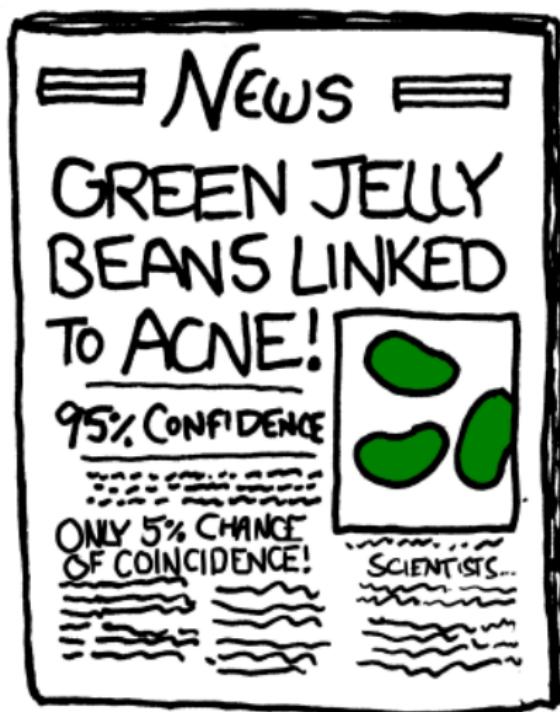
WE FOUND NO
LINK BETWEEN
YELLOW JELLY
BEANS AND ACNE
($P > 0.05$).



Publication Bias and NHST



Publication Bias and NHST



A “Tea-Test”

Consider the following example, given by Fisher in 1935:

In England, tea is often drunk together with milk. Suppose a tea expert claims to be able to tell whether the milk or the tea has been poured into a cup first. He is put to the test and is to state whether or not a given cup was produced by pouring milk first. His results are

correct, correct, correct, correct, correct, incorrect.

Question. What is the P -value of this test?

A “Tea-Test”

There are (at least) two interpretations of this example:

- ▶ The intention of the researcher, not just the raw experimental data, may determine the P -value of a test.
- ▶ In a hypothesis test, the outcomes that do **not** occur are just as important as the outcomes that do occur.

In particular, it is not considered to be good statistical practice to repeat an experiment to reject a null hypothesis until it is successful. To be probabilistically pure in the NHST sense, an experiment should be run once, and if the null hypothesis is not rejected, it should not be repeated.

Implications for Science

This causes problems of another nature - should an experiment that fails due to insufficient power really never be repeated ever again? That appears to be quite contrary to the nature of scientific inquiry.

Of course this is nonsense! In Fisher's approach, data may be obtained as often as desired, a test repeated as often as necessary, since the proof only serves as indirect evidence, not as a definitive rejection of the null hypothesis. In Neyman-Pearson, there are two alternatives and in a given situation, a decision is necessary. Therefore, data is gathered only once and a decision is made in the concrete circumstances. The fact that the alternative hypothesis is usually not just the negation of H_1 ensures that the result is meaningful.

However, since most researchers use the NHST approach, there is a large proportion of Type II errors in unpublished papers and many studies that would have led to good results that could not be obtained due to insufficient power (e.g., small sample size) are abandoned forever.

Single Sample Tests for the Mean and Variance

Instances of Hypothesis Tests

In this section, we will introduce various test statistics that can be used for either Fisher tests or Neyman-Pearson decision tests. In either case, the emphasis is first on rejecting some null hypothesis at a certain significance level, either directly in a Fisher test or by the test statistic being in a certain critical region. We will also discuss OC curves, as used in Neyman-Pearson tests for most of these tests.

We have already described how to perform tests for the mean based on the normal distribution with known variance (sometimes called **Z-tests**) and will not repeat these here.

The T -Test

17.1. **T -Test.** Let X_1, \dots, X_n be a random sample of size n from a normal distribution and let \bar{X} denote the sample mean, S^2 the sample variance. Let μ be the unknown population mean and μ_0 a null value of that mean. Then any test based on the statistic

$$T_{n-1} = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

is called a **T -test**.

We reject at significance level α

- ▶ $H_0: \mu = \mu_0$ if $|T_{n-1}| > t_{\alpha/2, n-1}$,
- ▶ $H_0: \mu \leq \mu_0$ if $T_{n-1} > t_{\alpha, n-1}$,
- ▶ $H_0: \mu \geq \mu_0$ if $T_{n-1} < -t_{\alpha, n-1}$.

The T -Test

17.2. Example. The breaking strength of a textile fiber is a normally distributed random variable. Specifications require that the mean breaking strength should equal 150 psi. The manufacturer would like to detect any significant departure from this value. Thus, he wishes to test

$$H_0: \mu = 150 \text{ psi} \quad H_1: |\mu - 150 \text{ psi}| > 2.5 \text{ psi}$$

A random sample of 15 fiber specimens is selected and their breaking strengths determined. The statistic

$$T = \frac{\bar{X} - \mu_0}{S/\sqrt{n}}$$

will follow a T_{14} -distribution. We specify $\alpha = 0.05$, and find $t_{0.025, 14} = 2.145$ from Table VI of the textbook. Thus, the critical region is given by $|t| > 2.145$.

The T -Test

The sample mean and variance are computed from the sample data as $\bar{x} = 152.18$ and $s^2 = 16.63$. Therefore, the test statistic is

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{152.18 - 150}{\sqrt{16.63/15}} = 2.07,$$

which does not fall into the critical region, so there is insufficient evidence to reject H_0 at the 5% level of significance.

Note that the T -distribution may be used for $\frac{\bar{X} - \mu_0}{S/\sqrt{n}}$ when a sample is obtained from a normal population. If a sample is obtained from a non-normal population, care must be taken; for large to medium sample sizes ($n \geq 25$) it can be shown that violating the normality assumption does not significantly change α and β . For small sample sizes, a T -test cannot be used and an alternative (non-parametric) test must be employed; such tests will be discussed later.

OC Curves for the T -Test

The OC curves for T -test have a similar appearance to those for the normal distribution. However, when calculating the probability of failing to reject H_0 if $\mu = \mu_0 + \delta$, $\delta > 0$, as we did for the normal distribution, a difficulty occurs. We obtain the quotient of a non-standardized ($\mu \neq 0$) normal distribution with a chi-distribution. This leads to the concept of ***non-central T-distributions***, which we will not go into here.

OC Curves for the T -Test

The OC curves for the T -distribution feature an abscissa whose scale is given by

$$d = \frac{|\mu - \mu_0|}{\sigma},$$

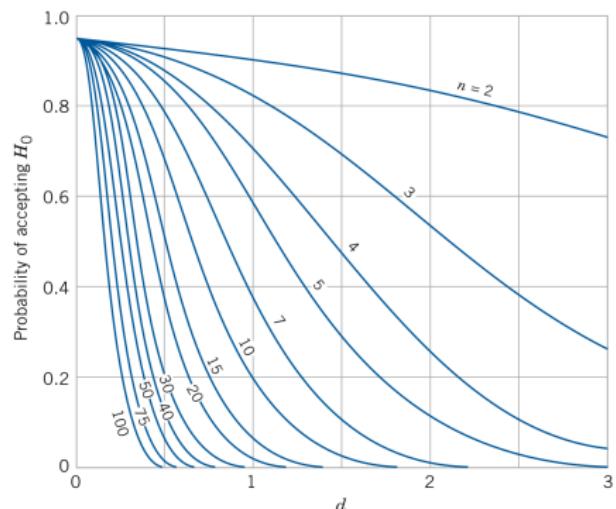
where σ is the ***unknown standard deviation*** of the random variable. We are left with three options:

1. If available, we can use prior experiments to insert a rough estimate for σ .
2. We can express the difference $\delta = |\mu - \mu_0|$ relative to σ , e.g., prescribing $d = \delta/\sigma < 1$ for a small difference in the mean or $d = \delta/\sigma < 2$ for a moderately large difference.
3. We substitute the sample standard deviation s for σ .

OC Curves for the T-Test

17.3. Example.

We return to Example 17.2. If the mean breaking strength of the fiber differs from 150 psi by 2.5 psi or more, we would like to reject the null hypothesis $H_0: \mu = 150$ psi with a probability of at least 0.9. Is the sample size $n = 15$ adequate to assure that the test is this sensitive?



If we use the previously obtained standard deviation $s = \sqrt{16.63} = 4.08$, then

$$d = \frac{|\mu - \mu_0|}{s} = \frac{2.5}{4.08} = 0.61.$$

The OC chart for $n = 15$, $\alpha = 0.05$, two-tailed, then gives $\beta \approx 0.45$. Thus the test is not powerful enough, since $1 - \beta = 0.55 < 0.9$.

The Chi-Squared Test

17.4. Chi-Squared Test. Let X_1, \dots, X_n be a random sample of size n from a normal distribution and let S^2 denote the sample variance. Let σ^2 be the unknown population variance and σ_0^2 a null value of that variance. Then a test for the variance based on the statistic

$$\chi_{n-1}^2 = \frac{(n-1)S^2}{\sigma_0^2}$$

is called a ***chi-squared test***. We reject at significance level α

- ▶ $H_0: \sigma = \sigma_0$ if $\chi_{n-1}^2 > \chi_{\alpha/2, n-1}^2$ or $\chi_{n-1}^2 < \chi_{1-\alpha/2, n-1}^2$,
- ▶ $H_0: \sigma \leq \sigma_0$ if $\chi_{n-1}^2 > \chi_{\alpha, n-1}^2$,
- ▶ $H_0: \sigma \geq \sigma_0$ if $\chi_{n-1}^2 < \chi_{1-\alpha, n-1}^2$.

The Chi-Squared Test

It is important to be aware of the following difficulty:

- ▶ The T -distribution can be used in the presence of large sample sizes for the distribution of the sample mean even if the underlying distribution is non-normal.
- ▶ It is, however, **not possible** to approximate the χ^2_{n-1} statistic in this way if the distribution is non-normal, regardless of sample size! Therefore, normality of the data must first be tested, and if the data is non-normal, other methods must be used.

The Chi-Squared Test

17.5. Example. One random variable studied while designing the front-wheel-drive half-shaft of a new model automobile is the displacement (in millimeters) of the constant velocity (CV) joints. With the joint angle fixed at 12° , 20 simulations were conducted, resulting in the following data:

6.2	1.9	4.4	4.9	3.5
4.6	4.2	1.1	1.3	4.8
4.1	3.7	2.5	3.7	4.2
1.4	2.6	1.5	3.9	3.2

For these data, $\bar{x} = 3.39$ and $s = 1.41$. Engineers designing the front-wheel-drive half-shaft claim that the standard deviation in the displacement of the CV shaft is less than 1.5 mm. Do these data support this contention?

The Chi-Squared Test

We can translate the described situation into a Fisher test for $H_0: \sigma^2 \geq 1.5$, which is equivalent to testing

$$H_0: \sigma^2 \geq 2.25.$$

From Table IV we obtain $\chi^2_{1-0.05, 19} = 10.1$. Hence the test will have a P -value of less than 0.05 if

$$\frac{(n-1)s^2}{\sigma_0^2} < 10.1.$$

The observed value of the test statistic is

$$\frac{19 \cdot 1.41^2}{2.25} = 16.79.$$

Since this value is greater than 10.1, there is no evidence to reject H_0 at the 5% level of significance.

OC Curves for the Chi-Squared Test

The abscissa parameter for the OC curves for the two-tailed chi-squared test is

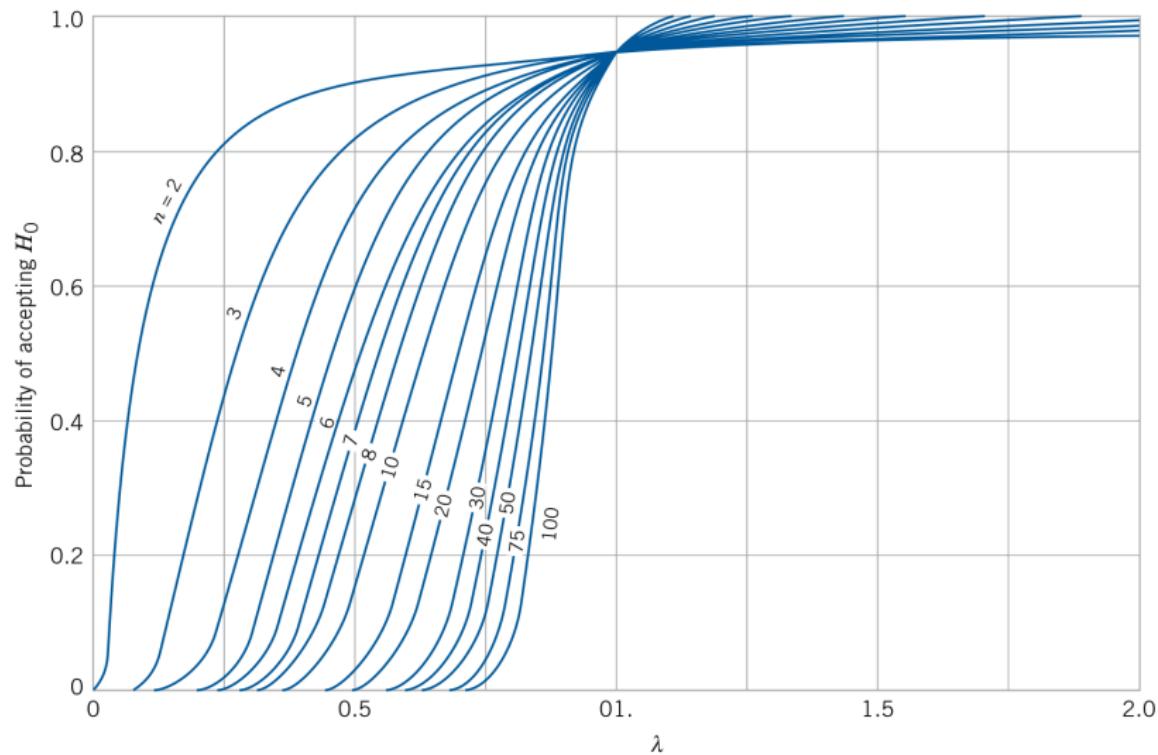
$$\lambda = \frac{\sigma}{\sigma_0}.$$

Note that the OC curves for the left- and right-tailed chi-squared distributions are distinct!

17.6. Example. Returning to Example 17.5, the engineers concerned are dissatisfied that H_0 was not rejected. A second test (this time of Neyman-Pearson type) is to be performed to establish that the standard deviation is less than $\sigma_0 = 1.5$ mm.

1. If we want to preset $\alpha = 0.05$, what is the critical region for the test at a sample size $n = 20$?
2. If $n = 20$, what true value of σ is necessary so that the test will have a power of $1 - \beta = 0.9$?
3. For $\alpha = 0.05$, make a statement on the sample size necessary to ensure that H_0 is rejected with 90% probability if $\sigma = 1.35$.

OC Curves for the Chi-Squared Test



OC Curves for Tests on the Variance

- From the table for the χ^2_{19} distribution we see that

$P[\chi^2_{1-\alpha, 19} \leq 10.1] = 0.05$, so the critical region for the variance is

$$\frac{(n-1)s^2}{\sigma_0^2} < 10.1 \quad \Leftrightarrow \quad s^2 < \frac{2.25 \cdot 10.1}{19} = 1.20$$

i.e., $s < 1.09$.

- For $n = 20$, the line intersects the horizontal rule $\beta = 0.1$ at $\lambda = 0.6$.
This means that

$$\sigma < 0.6\sigma_0 = 0.9$$

is necessary for H_0 to be rejected 90% of the time.

- The graph shows that a sample size significantly larger than $n = 100$ would be necessary.

Non-Parametric Single Sample Tests for the Median

Non-Parametric Statistics

Previously: used methods based on **normal distribution**

Now: methods that work more generally, without any assumption on the random variable X .

Two basic concepts:

- **non-parametric statistics** do not assume the dependence on any parameter.

18.1. Example. The confidence interval for the mean derived previously has the form

$$\bar{X} \pm z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \quad \text{or} \quad \bar{X} \pm t_{\alpha/2, n-1} \frac{s}{\sqrt{n}},$$

which uses the parameters $z_{\alpha/2}$ and σ (or $t_{\alpha/2, n-1}$).

In contrast, in the assignments we have studied a non-parametric confidence interval for the median which does not use any parameter that is not directly derived from the random sample.

Non-Parametric Statistics

Two basic concepts:

- ▶ ***non-parametric statistics*** do not assume the dependence on any parameter.
- ▶ ***distribution-free statistics*** do not assume that X follows any particular distribution (such as the normal distribution).

Although different, both types of methods are loosely referred to as ***non-parametric methods***.

Generally, one uses

- ▶ the ***median or other location measure*** instead of the mean;
- ▶ the ***interquartile range or other dispersion measure*** instead of the variance.

Sign Test for the Median

Recall that the median of a random variable X is defined as the value M such that

$$P[X \leq M] = P[X \geq M] = 1/2.$$

The **sign test** will have a null hypothesis of either the two-tailed or one-tailed form

- ▶ $H_0: M = M_0$
- ▶ $H_0: M \leq M_0$ or $H_0: M \geq M_0$

and is usually implemented as a **Fisher test**.

Sign Test for the Median

The idea is simple: Given a random sample X_1, \dots, X_n of size n from X , each measurement has a $1/2$ probability of being smaller than M and a $1/2$ probability of being larger than M .

(We neglect for now the possibility of $X_k = M$.)

If significantly less than one-half of the sample measurements is less than or greater than M_0 , this may be taken as evidence to reject H_0 .

Given a sample X_1, \dots, X_n , define

$$Q_+ = \#\{X_k : X_k - M_0 > 0\}, \quad Q_- = \#\{X_k : X_k - M_0 < 0\}.$$

So Q_+ is the number of “positive signs” and Q_- the number of “negative signs.” We note that

$$P[Q_- \leq k \mid M = M_0] = \sum_{x=0}^k \binom{n}{x} \frac{1}{2^n}$$

Sign Test for the Median

18.2. Sign Test. Let X_1, \dots, X_n be a random sample of size n from an arbitrary continuous distribution and let

$$Q_+ = \#\{X_k : X_k - M_0 > 0\}, \quad Q_- = \#\{X_k : X_k - M_0 < 0\}.$$

We reject at significance level α

- ▶ $H_0: M \leq M_0$ if $P[Q_- < k \mid M = M_0] < \alpha$,
- ▶ $H_0: M \geq M_0$ if $P[Q_+ < k \mid M = M_0] < \alpha$,
- ▶ $H_0: M = M_0$ if $P[\min(Q_-, Q_+) < k \mid M = M_0] < \alpha/2$.

Sign Test for the Median

18.3. Example. A certain six-sided die is suspected of being unbalanced. Based on past experience, it is suspected that the median is greater than 3.5. We decide to test the null hypothesis

$$H_0: M \leq 3.5.$$

The die is rolled 20 times, yielding the following results:

X_i	$X_i - M_0$	Sign	X_i	$X_i - M_0$	Sign	X_i	$X_i - M_0$	Sign
5	1.5	+	3	-0.5	-	4	0.5	+
1	-2.5	-	6	2.5	+	4	0.5	+
5	1.5	+	2	-1.5	-	4	0.5	+
4	0.5	+	3	-0.5	-	3	-0.5	-
4	0.5	+	5	1.5	+	3	-0.5	-
6	2.5	+	5	1.5	+	4	0.5	+
6	2.5	+	6	2.5	+			

Sign Test for the Median

We note that there are 6 negative signs,

$$Q_- = 6.$$

We then find that

$$P[Q_- \leq 6 \mid M = 3.5] = \frac{1}{2^{20}} \sum_{x=0}^6 \binom{20}{x} = 0.0577.$$

This is the P -value of the test. It would be reasonable to decide not to reject H_0 , i.e., the results do not provide convincing evidence that H_0 is false.

Assumptions, Limitations and Issues

Advantages:

- ▶ Very flexible, no assumptions on distribution of X .
- ▶ Magnitude of $X_i - M_0$ is not needed.

Disadvantages:

- ▶ Not very powerful.

Possible Issues:

- ▶ In some situations, especially when sampling from a discrete distribution, it may happen that

$$X_i - M_0 = 0.$$

In such case, usual practice is to ***exclude the data from the analysis.***

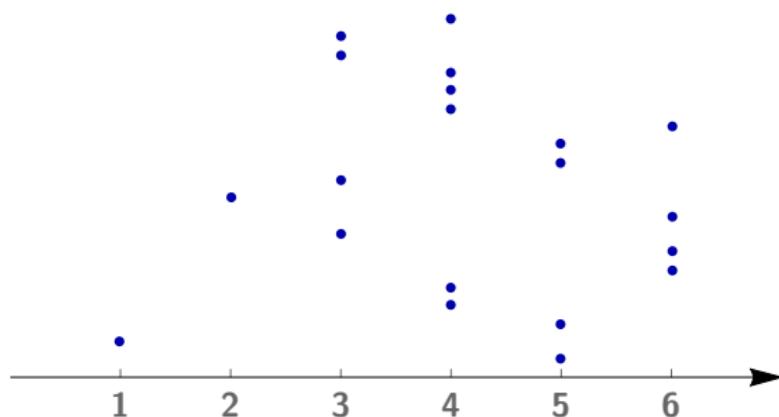
Wilcoxon Signed Rank Test

The power of the sign test can be increased by taking the magnitude of $X_i - M_0$ into account.

In order to avoid using parameters, Wilcoxon introduced the notion of **ranks**: observations are ranked from smallest to largest and instead of considering simply their sign, one analyzes the **signed rank**.



Frank Wilcoxon (1892-1965). A R Sampson and B Spencer, A conversation with L. Richard Savage, Statistical Science 14 (1999), 126-148.



Wilcoxon Signed Rank Test

This analysis of ranks supposes that the data comes from a distribution that is ***symmetric about its median***. This assumption was not needed for the sign test.

The data is ranked from smallest to largest absolute difference to the null value of the median. In other words, the observation where $|X_i - M_0|$ is smallest will be ranked first and be assigned rank $R_i = 1$, while the observation where $|X_j - M_0|$ is largest will receive rank $R_j = n$.

The signed rank is found by multiplying the rank with -1 if $X_i - M_0 < 0$ and $+1$ if $X_i - M_0 > 1$.

The positive ranks as well as the negative ranks are summed separately, yielding two statistics W_+ and W_- .

Ties in ranks are assigned the ***average of their ranks***. Hence, the total sum of the ranks is always $n(n + 1)/2$.

Test Tables and Normal Approximation

The distribution of the test statistics is complicated; there are tables that give critical values for small sample sizes, typically up to $n \leq 20$.

For non-small sample sizes ($n \geq 10$) a normal distribution with parameters

$$\text{E}[W] = \frac{n(n+1)}{4}, \quad \text{Var}[W] = \frac{n(n+1)(2n+1)}{24}.$$

may be used as an approximation. However, in that case the variance needs to be reduced if there are ties: for each group of t ties, the variance is reduced by $(t^3 - t)/48$.

Wilcoxon Signed Rank Test

18.4. Wilcoxon Signed Rank Test. Let X_1, \dots, X_n be a random sample of size n from a symmetric distribution. Order the n absolute differences $|X_i - M|$ according to magnitude, so that $X_{R_i} - M_0$ is the R_i th smallest difference by modulus. If ties in the rank occur, the mean of the ranks is assigned to all equal values.

Let

$$W_+ = \sum_{R_i > 0} R_i, \quad |W_-| = \sum_{R_i < 0} |R_i|.$$

We reject at significance level α

- ▶ $H_0: M \leq M_0$ if W_- is smaller than the critical value for α ,
- ▶ $H_0: M \geq M_0$ if W_+ is smaller than the critical value for α ,
- ▶ $H_0: M = M_0$ if $W = \min(W_+, |W_-|)$ is smaller than the critical value for $\alpha/2$.

Ranking the Results of the Die Rolls

18.5. Example. Returning to the previous Example 18.3, we want to test $H_0: M \leq 3.5$ and have the following observations, ordered from smallest to largest:

X_i	$X_i - M_0$	R_i	X_i	$X_i - M_0$	R_i
3	-0.5	-5.5	2	-1.5	-13
3	-0.5	-5.5	5	1.5	+13
3	-0.5	-5.5	5	1.5	+13
3	-0.5	-5.5	5	1.5	+13
4	0.5	+5.5	5	1.5	+13
4	0.5	+5.5	1	-2.5	-18
4	0.5	+5.5	6	2.5	+18
4	0.5	+5.5	6	2.5	+18
4	0.5	+5.5	6	2.5	+18
4	0.5	+5.5	6	2.5	+18

Finding the P -Value

We calculate the sum of the negative ranks,

$$w_- = -5.5 - 5.5 - 5.5 - 5.5 - 13 - 18 = -53.$$

Consulting a table, the critical value for $n = 20$ and $\alpha = 0.05$ is 60. For $\alpha = 0.01$ it is 43. Since $|w_-|$ lies between these values, the P -value of the test is between 1% and than 5%, most likely around 2%-3%.

Alternatively, we may use the normal distribution with mean $\mu = n(n + 1)/4 = 105$ and variance

$$\sigma^2 = \frac{n(n + 1)(2n + 1)}{24} - \frac{10^3 - 10}{48} - 2 \cdot \frac{5^3 - 5}{48}.$$

Then

$$z = \frac{|w_-| - \mu}{\sigma} = -1.977$$

and we find that $P[Z < -1.977] = 0.024$.

Conclusion of the Test

It may reasonable to reject H_0 based on this P -value. There is some evidence that the die does not follow a symmetric distribution with median less than or equal to 3.5.

In practice, we may come to several conclusions:

- ▶ the die results follow a non-symmetric distribution; or
- ▶ the die results follow a symmetric distribution, but the median is greater than 3.5.

This example features many ties between results. In general, the power if the signed rank test is reduced if there are very many ties and some modified tests have been recently proposed to improve this.

Nevertheless, we obtained a result with a smaller P -value than we did for the Sign Test.

Assumptions, Limitations and Issues

Advantages:

- ▶ Fairly powerful; may even be used as an alternative to the T -test without much loss of power.

Disadvantages:

- ▶ Assumes a symmetric distribution around the median.

Possible Issues:

- ▶ As in the sign test, observations where $X_i - M_0 = 0$ are discarded.
- ▶ Some authors prefer to use a modified but equivalent version of the test, where all positive and negative ranks are added together. The test is equivalent, but different tables need to be used.

Hypothesis Tests with Mathematica

We can use Mathematica for calculating test statistics. Suppose we have the following data:

```
data := {41.50, 41.38, 42.24, 41.85, 41.76,  
        42.08, 41.62, 42.16, 41.71, 41.44}
```

We want to test $H_0: \mu \leq \mu_0 = 41.5$, assuming a known variance of $\sigma^2 = 0.1$. The Z-test statistic is

```
 $\bar{x} := \text{Mean}[data]; n := \text{Length}[data]; \sigma_0 := \sqrt{0.1}; \mu_0 := 41.5;$ 
```

$$Z = \frac{\bar{x} - \mu_0}{\sigma_0 / \sqrt{n}}$$

2.74

We can then find a P -value for the test:

```
1 - CDF[NormalDistribution[0, 1], z]
```

0.00307196

Hypothesis Tests with Mathematica

Mathematica also has many standard tests built-in. The previous test can be performed as follows:

```
Needs["HypothesisTesting`"]

ZTest[data, 0.1, 41.5, "TestDataTable",
  AlternativeHypothesis -> "Greater"]
```

	Statistic	P-Value
Z	2.74	0.00307196

The corresponding two-tailed test $H_0: \mu = \mu_0$ would yield:

```
ZTest[data, 0.1, 41.5, "TestDataTable",
  AlternativeHypothesis -> "Unequal"]
```

	Statistic	P-Value
Z	2.74	0.00614392

Hypothesis Tests with Mathematica

Of course, there are also T-tests:

```
TTest[data, 41.5, "TestDataTable",
  AlternativeHypothesis -> "Unequal"]
```

	Statistic	P-Value
T	2.8439	0.0192801

The sign test and the Wilcoxon signed rank test are also implemented:

```
SignTest[data, 41.5, "TestDataTable",
  AlternativeHypothesis -> "Unequal"]
```

	Statistic	P-Value
Sign	7	0.179687

```
SignedRankTest[data, 41.5, "TestDataTable",
  AlternativeHypothesis -> "Unequal"]
```

	Statistic	P-Value
Signed-Rank	41.5	0.028263

Hypothesis Tests with Mathematica

The chi-squared test is called the Fisher ratio test in Mathematica:

```
FisherRatioTest[data, 0.1, "TestDataTable",
  AlternativeHypothesis -> "Unequal"]
```

	Statistic	P-Value
Fisher Ratio	8.3544	0.997726

We can verify the result by hand:

$$x := \sqrt{\frac{(n-1) \text{ Variance}[data]}{\sigma_0^2}} ; \chi^2$$

8.3544

$2 (1 - \text{CDF}[\text{ChiSquareDistribution}[n-1], x^2])$

0.997726

Note the behavior of the cumulative distribution function for the chi-squared distribution and the doubling of the *P*-value!

Inferences on Proportions

Estimating Proportions

One of the (mathematically) simplest population parameters of general interest is the **proportion** of members of a population with some trait. Every member of the population is characterized as either having or not having this trait. We describe this mathematically by defining the random variable

$$X = \begin{cases} 1 & \text{has trait,} \\ 0 & \text{does not have trait.} \end{cases}$$

The proportion of the members of the population having the trait is

$$p = \frac{\# \text{ members wih trait}}{\text{population size}} = \frac{1}{N} \sum_{i=1}^N x_i$$

where N is the population size and x_i is the value of the variable X for the i th member of the population. Hence the proportion is equal to the mean of X .

Estimating Proportions

It follows that if we take a random sample X_1, \dots, X_n of X , the sample mean

$$\hat{p} = \bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$$

is an (unbiased) estimator for p .

The random variable X follows a Bernoulli distribution with expectation $E[X] = p$ and variance $\text{Var } X = p(1 - p)$. However, it is often more convenient to use a normal distribution for confidence intervals and hypothesis tests.

By the central limit theorem, \hat{p} is approximately normally distributed with mean p and variance $p(1 - p)/n$. Hence,

$$\frac{\hat{p} - p}{\sqrt{p(1 - p)/n}}$$

is approximately standard-normally distributed.

Estimating Proportions

It follows immediately that the following is a $100(1 - \alpha)\%$ confidence interval for p :

$$\hat{p} \pm z_{\alpha/2} \sqrt{p(1 - p)/n}$$

But the interval depends on the unknown parameter p , which we are actually trying to estimate! One solution to the problem is to replace p by \hat{p} , i.e., to write

$$\hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}.$$

But then the number $z_{\alpha/2}$ is no longer accurate (when we replaced σ by S to obtain a confidence interval for the mean, we had to switch from $z_{\alpha/2}$ to $t_{\alpha/2}$).

However, we are approximating the binomial distribution in any case - we might argue that if the sample size n is large enough to allow the central limit theorem to hold, then the difference between $z_{\alpha/2}$ and a corrected value will be negligible. This is not a perfect solution, but a detailed discussion would lead to far here.

Estimating Proportions

19.1. Example. In 2017, the Institute of Sociology of the Shanghai Academy of Social Sciences conducted a survey among residents of Shanghai to ask their opinion about the municipal proposal to limit the numbers of residents to 25 million by 2020 and keep that number stable until 2040.

Among the 2079 residents surveyed, 48.5% indicated that this measure would benefit Shanghai's development. Assuming that those questioned constitute a random sample of Shanghai residents, a 99% confidence interval for the proportion of residents with this opinion is given by

$$p = 0.485 \pm 2.575 \sqrt{0.485 \cdot 0.515 / 2079} = 0.485 \pm 0.028$$

Literature: [https://archive.shine.cn/metro/society/
Pros-and-cons-of-limiting-citys-population/shdaily.shtml](https://archive.shine.cn/metro/society/Pros-and-cons-of-limiting-citys-population/shdaily.shtml)

Choosing the Sample Size

As a practical matter, we are often able to choose (perhaps within constraints) the sample size. We may want to be able to claim that “with $xx\%$ probability, \hat{p} differs from p by at most d .”

Given a $100(1 - \alpha)\%$ confidence interval $p = \hat{p} \pm z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}$, we know with $100(1 - \alpha)\%$ confidence that

$$d = z_{\alpha/2} \sqrt{\hat{p}(1 - \hat{p})/n}.$$

Given d , this means that we should choose

$$n = \frac{z_{\alpha/2}^2 \hat{p}(1 - \hat{p})}{d^2}$$

to ensure that $|p - \hat{p}| < d$ with $100(1 - \alpha)\%$ confidence. However, this formula requires us to have an estimate \hat{p} of p beforehand.

Choosing the Sample Size

If no estimate for p is available, we can at least use that $x(1 - x) < 1/4$ for all $x \in \mathbb{R}$ to deduce that

$$n = \frac{z_{\alpha/2}^2}{4d^2}$$

will ensure $|p - \hat{p}| < d$ with $100(1 - \alpha)\%$ confidence.

19.2. Example. How large a sample is needed to estimate the proportion of members in a population with a certain trait to within 0.02 with 90% confidence?

Since no prior estimate is available, we take

$$n = \frac{z_{0.05}^2}{4d^2} = \frac{1.645^2}{4 \cdot 0.02^2} = 1692.$$

Hypothesis Testing

19.3. Test for Proportion. Let X_1, \dots, X_n be a random sample of (large) size n from a Bernoulli distribution with parameter p and let $\hat{p} = \bar{X}$ denote the sample mean. Then any test based on the statistic

$$Z = \frac{\hat{p} - p_0}{\sqrt{p_0(1 - p_0)/n}}$$

is called a **large-sample test for proportion**.

We reject at significance level α

- ▶ $H_0: p = p_0$ if $|Z| > z_{\alpha/2}$,
- ▶ $H_0: p \leq p_0$ if $Z > z_{\alpha}$,
- ▶ $H_0: p \geq p_0$ if $Z < -z_{\alpha}$.

Comparing Two Proportions

Two Populations:

- ▶ $X^{(1)} \sim \text{Bernoulli}(p_1)$,
- ▶ $X^{(2)} \sim \text{Bernoulli}(p_2)$.

Goal: make inferences on $p_1 - p_2$.

Suppose a random sample of size n_1 from population 1 and another random sample of size n_2 from population 2 are given.

An unbiased estimator for $p_1 - p_2$ is

$$\widehat{p_1 - p_2} := \widehat{p}_1 - \widehat{p}_2 = \overline{X}^{(1)} - \overline{X}^{(2)},$$

where $\overline{X}^{(1)}$ and $\overline{X}^{(2)}$ are the sample means of the respective random samples.

A Confidence Interval

We have the approximate distributions

$$\bar{X}^{(1)} \sim N\left(p_1, \frac{p_1(1-p_1)}{n_1}\right), \quad \bar{X}^{(2)} \sim N\left(p_2, \frac{p_2(1-p_2)}{n_2}\right)$$

We can infer that for large samples

$$\widehat{p_1 - p_2} \sim N\left(p_1 - p_2, \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}\right)$$

This allows us to deduce the following $100(1 - \alpha)\%$ confidence interval for $p_1 - p_2$:

$$\widehat{p_1} - \widehat{p_2} \pm z_{\alpha/2} \sqrt{\frac{\widehat{p}_1(1 - \widehat{p}_1)}{n_1} + \frac{\widehat{p}_2(1 - \widehat{p}_2)}{n_2}}$$

which is valid for large sample sizes.

Comparing Two Proportions

19.4. Test for Comparing Two Proportions. Suppose two random samples of (large) sizes n_1 and n_2 from two Bernoulli distributions with parameters p_1 and p_2 are given. Denote by \hat{p}_1 and \hat{p}_2 the means of the two samples.

Let $(p_1 - p_2)_0$ be a null value for the difference $p_1 - p_2$. Then the test based on the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2 - (p_1 - p_2)_0}{\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}}$$

is called a **large-sample test for differences in proportions**.

We reject at significance level α

- ▶ $H_0: p_1 - p_2 = (p_1 - p_2)_0$ if $|Z| > z_{\alpha/2}$,
- ▶ $H_0: p_1 - p_2 \leq (p_1 - p_2)_0$ if $Z > z_\alpha$,
- ▶ $H_0: p_1 - p_2 \geq (p_1 - p_2)_0$ if $Z < -z_\alpha$.

Pooled Estimator for the Proportion

Most commonly we test against the null value $(p_1 - p_2)_0 = 0$, i.e.,

$$H_0: p_1 = p_2.$$

If H_0 is true, the common proportion is

$$p = p_1 = p_2.$$

Then both \hat{p}_1 and \hat{p}_2 are estimators for p .

It turns out that the best course of action is to take the weighted average:
we define the **pooled estimator for the proportion**,

$$\hat{p} := \frac{n_1 \hat{p}_1 + n_2 \hat{p}_2}{n_1 + n_2}. \quad (19.1)$$

Pooled Test for Equality of Proportions

19.5. Pooled Test for Equality of Proportions. Suppose two samples of (large) sizes n_1 and n_2 from two Bernoulli distributions with parameters p_1 and p_2 are given. Denote by \hat{p}_1 and \hat{p}_2 the means of the two samples. Let \hat{p} be the pooled estimator for the proportion. Then the test based on the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}}.$$

is called a *pooled large-sample test for equality of proportions*.

We reject at significance level α

- ▶ $H_0: p_1 = p_2$ if $|Z| > z_{\alpha/2}$,
- ▶ $H_0: p_1 \leq p_2$ if $Z > z_\alpha$,
- ▶ $H_0: p_1 \geq p_2$ if $Z < -z_\alpha$.

Pooled Proportions

19.6. Example. Many consumers think that automobiles built on Mondays are more likely to have serious defects than those built on any other day of the week. To support this theory, a random sample of 100 cars built on Monday is selected and inspected. Of these, eight are found to have serious defects. A random sample of 200 cars produced on other days reveals 12 with serious defects. Do these data support the stated contention?

We test

$$H_0: p_1 \leq p_2.$$

where p_1 denotes the proportion of cars with serious defects produced on Mondays.

Estimates for p_1 and p_2 are

$$\hat{p}_1 = 8/100 = 0.08, \quad \hat{p}_2 = 12/200 = 0.06.$$

Pooled Proportions

The pooled estimate for the common population proportion is

$$\hat{p} = \frac{100 \cdot 0.08 + 200 \cdot 0.06}{100 + 200} = 20/300 = 0.066.$$

The observed value of the test statistic is

$$\frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1 - \hat{p}) \left(\frac{1}{n_1} + \frac{1}{n_2} \right)}} = \frac{0.08 - 0.06}{\sqrt{0.066 \cdot 0.934 \left(\frac{1}{100} + \frac{1}{200} \right)}} = 0.658.$$

From the standard normal table, we see that the probability of observing this large or a larger value is 0.2546, so there is no evidence that H_0 might be false.

Comparing Two Variances

Two Normally-Distributed Populations:

- ▶ $X^{(1)} \sim N(\mu_1, \sigma_1^2)$,
- ▶ $X^{(2)} \sim N(\mu_2, \sigma_2^2)$.

Goal: Develop a test to compare σ_1^2 and σ_2^2 .

Taking samples of sizes n_1 and n_2 from the populations, we know that

$$\frac{(n_1 - 1)S_1^2}{\sigma_1^2} \sim \chi_{n_1-1}^2, \quad \frac{(n_2 - 1)S_2^2}{\sigma_2^2} \sim \chi_{n_2-1}^2.$$

The difference of two chi-squared distributions is hard to analyze, so we may be better off looking at the quotient:

$$\sigma_1^2 = \sigma_2^2$$

if and only if

$$\frac{\sigma_1^2}{\sigma_2^2} = 1$$

The F -Distribution

20.1. Definition. Let $X_{\gamma_1}^2$ and $X_{\gamma_2}^2$ be independent chi-squared random variables with γ_1 and γ_2 degrees of freedom, respectively.

The random variable

$$F_{\gamma_1, \gamma_2} = \frac{X_{\gamma_1}^2 / \gamma_1}{X_{\gamma_2}^2 / \gamma_2}$$

is said to follow an **F -distribution with γ_1 and γ_2 degrees of freedom**.

20.2. Remark. From the definition, it is clear that

$$P[F_{\gamma_1, \gamma_2} < x] = P\left[\frac{1}{F_{\gamma_1, \gamma_2}} > \frac{1}{x}\right] = 1 - P\left[F_{\gamma_2, \gamma_1} < \frac{1}{x}\right],$$

so the density functions of the F_{γ_1, γ_2} and F_{γ_2, γ_1} -distributions are related.

Density of the F -Distribution

Using Theorem 10.2 for the density of the quotient of two independent variables we can calculate the density function explicitly:

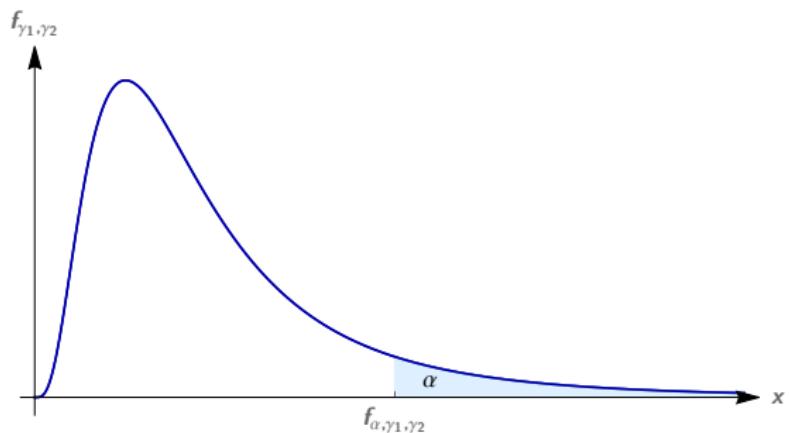
20.3. Lemma. The density of a random variable following an F -distribution with γ_1 and γ_2 degrees of freedom is given by

$$f_{\gamma_1, \gamma_2}(x) = \gamma_1^{\gamma_1/2} \gamma_2^{\gamma_2/2} \frac{\Gamma(\frac{\gamma_1 + \gamma_2}{2})}{\Gamma(\frac{\gamma_1}{2}) \Gamma(\frac{\gamma_2}{2})} \frac{x^{\gamma_1/2 - 1}}{(\gamma_1 x + \gamma_2)^{(\gamma_1 + \gamma_2)/2}}$$

for $x \geq 0$ and $f_{\gamma_1, \gamma_2}(x) = 0$ for $x < 0$.

Critical Points of the F -Distribution

For $0 < \alpha < 1$, we define the point $f_{\alpha, \gamma_1, \gamma_2}$ by $P[F_{\gamma_1, \gamma_2} > f_{\alpha, \gamma_1, \gamma_2}] = \alpha$.



Selected critical values $f_{\alpha, \gamma_1, \gamma_2}$ are tabulated. Since the F distribution has two parameters γ_1 and γ_2 , often only the values $f_{0.1, \gamma_1, \gamma_2}$ and $f_{0.05, \gamma_1, \gamma_2}$ are listed for various values of γ_1 and γ_2 .

Critical Points of the F -Distribution

From Remark 20.2 we see that

$$\begin{aligned}1 - \alpha &= P[F_{\gamma_1, \gamma_2} \geq f_{1-\alpha, \gamma_1, \gamma_2}] \\&= 1 - P[F_{\gamma_1, \gamma_2} < f_{1-\alpha, \gamma_1, \gamma_2}] \\&= P[F_{\gamma_2, \gamma_1} < 1/f_{1-\alpha, \gamma_1, \gamma_2}] \\&= 1 - P[F_{\gamma_2, \gamma_1} \geq 1/f_{1-\alpha, \gamma_1, \gamma_2}] \\&\stackrel{!}{=} 1 - P[F_{\gamma_2, \gamma_1} \geq f_{\alpha, \gamma_2, \gamma_1}]\end{aligned}$$

so

$$f_{1-\alpha, \gamma_1, \gamma_2} = \frac{1}{f_{\alpha, \gamma_2, \gamma_1}}. \quad (20.1)$$

It follows that the values of the “right-tail” critical point $f_{\alpha, \gamma_1, \gamma_2}$ are sufficient to find corresponding “left-tail” points.

The F -Distribution

20.4. Theorem. Let S_1^2 and S_2^2 be sample variances based on independent random samples of sizes n_1 and n_2 drawn from normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively.

If $\sigma_1^2 = \sigma_2^2$, then the statistic

$$S_1^2/S_2^2$$

follows an F -distribution with $n_1 - 1$ and $n_2 - 1$ degrees of freedom.

Proof.

We know that $(n_1 - 1)S_1^2/\sigma_1^2$ and $(n_2 - 1)S_2^2/\sigma_2^2$ follow chi-squared distributions with $n_1 - 1$ and $n_2 - 1$ degrees of freedom, respectively. Then

$$F_{n_1-1, n_2-1} = \frac{[(n_1 - 1)S_1^2/\sigma_1^2]/(n_1 - 1)}{[(n_2 - 1)S_2^2/\sigma_2^2]/(n_2 - 1)} = \frac{\sigma_2^2 S_1^2}{\sigma_1^2 S_2^2}.$$

If $\sigma_1^2 = \sigma_2^2$, this reduces to S_1^2/S_2^2 . □

The F -Test

20.5. F -Test. Let S_1^2 and S_2^2 be sample variances based on independent random samples of sizes n_1 and n_2 drawn from normal populations with means μ_1 and μ_2 and variances σ_1^2 and σ_2^2 , respectively. Then a test based on the statistic

$$F_{n_1-1, n_2-1} = \frac{S_1^2}{S_2^2}$$

is called an **F -test**.

We reject at significance level α

- ▶ $H_0: \sigma_1 \leq \sigma_2$ if $\frac{S_1^2}{S_2^2} > f_{\alpha, n_1-1, n_2-1}$,
- ▶ $H_0: \sigma_1 \geq \sigma_2$ if $\frac{S_2^2}{S_1^2} > f_{\alpha, n_2-1, n_1-1}$,
- ▶ $H_0: \sigma_1 = \sigma_2$ if $\frac{S_1^2}{S_2^2} > f_{\alpha/2, n_1-1, n_2-1}$ or $\frac{S_2^2}{S_1^2} > f_{\alpha/2, n_2-1, n_1-1}$

Remarks on the F -Test

20.6. Remarks.

- ▶ We have used (20.1) and written the critical regions in terms of right-tailed points; note the subscripts of the critical points carefully!
- ▶ For the F -test to be applicable, it is essential that the populations are normally distributed.
- ▶ If possible, the sample sizes n_1 and n_2 should be equal.
- ▶ It turns out that the F -test is not very powerful; β can be quite large. In order to keep β small, one often tests at $\alpha = 0.1$ or $\alpha = 0.2$ level of significance.
- ▶ When testing to see whether two population variances are equal for the purposes of later applying other tests, such as a comparison of their means, one ***hopes to not reject H_0*** ! In that case, the probability of committing a (Type II) error is given by β and a small β is more important than a small α .

Comparing Two Variances - The F -Test

20.7. Example. Chemical etching is used to remove copper from printed circuit boards. X_1 and X_2 represent process yields in % when two different concentrations are used. Suppose that we wish to test

$$H_0: \sigma_1^2 = \sigma_2^2.$$

Two samples of sizes $n_1 = n_2 = 8$ yield $s_1^2 = 4.02$ and $s_2^2 = 3.89$, and

$$\frac{s_1^2}{s_2^2} = \frac{4.02}{3.89} = 1.03.$$

From Table IX we see that $f_{0.1;7,7} = 2.785$. Since our test statistic is much smaller than this value, the P -value of the (two-tailed) test is significantly greater than $2 \cdot 0.1 = 20\%$. There is not enough evidence to reject H_0 .

OC Curves for the *F*-Test

For the case $n_1 = n_2 = n$ there are OC curves plotting β against the parameter

$$\lambda = \frac{\sigma_1}{\sigma_2}.$$

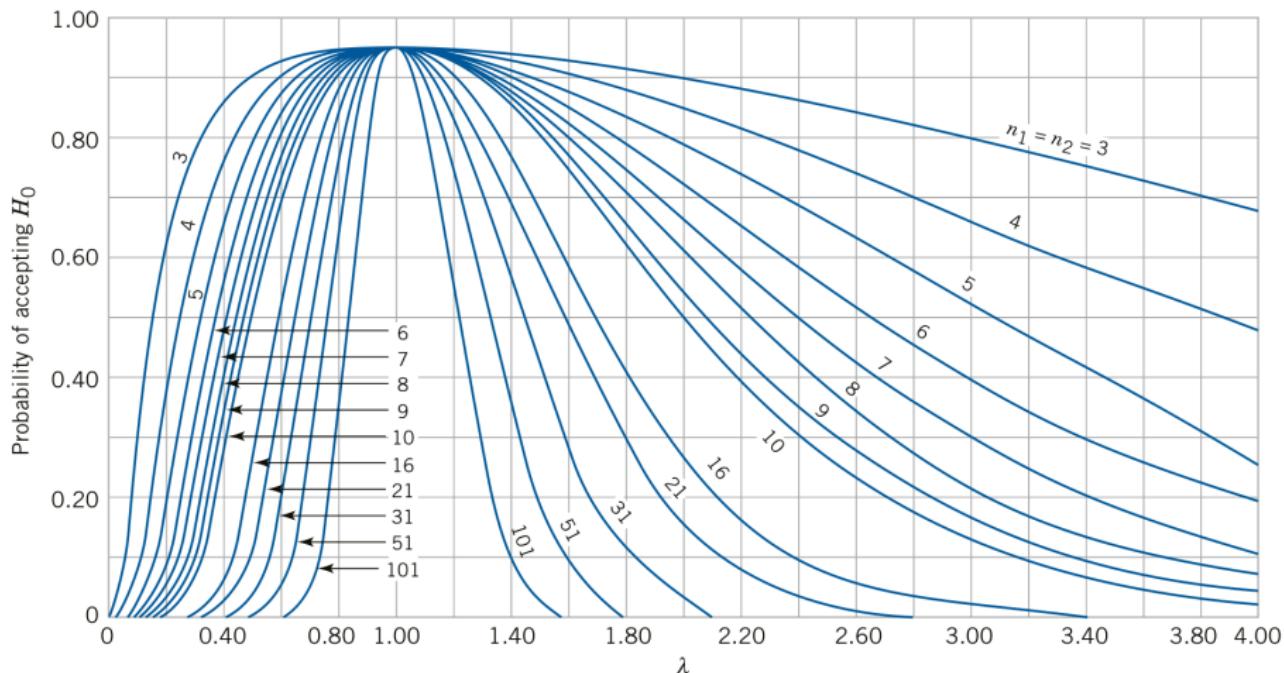
20.8. Example. Continuing from Example 20.7, suppose that one of the concentrations affected the variance of the yield so that one of the variances was four times the other and we wished to detect this with probability at least 0.80. What sample size should be used?

For this situation, a Neyman-Pearson test should be used:

$$H_0: \sigma_1^2 = \sigma_2^2, \quad H_1: \max\left(\frac{\sigma_1^2}{\sigma_2^2}, \frac{\sigma_1^2}{\sigma_2^2}\right) \geq 4$$

If one variance is four times the other, then $\lambda = \sigma_1/\sigma_2 = 2$.

OC Curves for the F -Test



From the OC chart, we see that a sample size of about 20 will be sufficient.

Comparison of Two Means

Comparing Two Means

Two Normally-Distributed Populations:

- ▶ $X^{(1)} \sim N(\mu_1, \sigma_1^2),$
- ▶ $X^{(2)} \sim N(\mu_2, \sigma_2^2).$

Goal: compare μ_1 and μ_2 .

Three Basic Cases:

- ▶ σ_1^2 and σ_2^2 are known
- ▶ σ_1^2 and σ_2^2 are unknown but $\sigma_1^2 = \sigma_2^2$
- ▶ σ_1^2 and σ_2^2 are unknown and not necessarily equal

Also:

- ▶ paired comparisons
- ▶ non-parametric tests

A Point Estimator for the Difference of Means

We take random samples $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ of sizes n_1 and n_2 from the populations, we can find a point estimator for the difference of the two means

$$\widehat{\mu_1 - \mu_2} := \widehat{\mu}_1 - \widehat{\mu}_2 = \bar{X}^{(1)} - \bar{X}^{(2)}.$$

Since

$$\bar{X}^{(1)} \sim N(\mu_1, \sigma_1^2/n_1), \quad \bar{X}^{(2)} \sim N(\mu_2, \sigma_2^2/n_2),$$

we see that $\bar{X}_1 - \bar{X}_2$ is normal with mean $\mu_1 - \mu_2$ and variance $\sigma_1^2/n_1 + \sigma_2^2/n_2$, i.e.,

$$\frac{\bar{X}^{(1)} - \bar{X}^{(2)} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

is a standard normal random variable.

Neyman-Pearson Test with Variances Known

We may use this result to obtain confidence intervals for the difference of means and to conduct hypothesis tests.

21.1. Example. The plant manager at an orange juice canning facility is interested in comparing the performance of two different production lines in her plant. As line number 1 is relatively new, she suspects that its output in number of cases per day is greater than the number of cases produced by the older line 2.

She sets up the hypotheses

$$H_0: \mu_1 \leq \mu_2, \quad H_1: \mu_1 > \mu_2 + 10 \text{ cases}.$$

She decides to use $\alpha = 5\%$. The test statistic is

$$Z = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

and H_0 will be rejected if this number is greater than $z_{0.05} = 1.645$.

Neyman-Pearson Test with Variances Known

From experience with operating this type of equipment it is known that $\sigma_1^2 = 40$ and $\sigma_2^2 = 50$.

Ten days of data are selected at random for each line, for which it is found that $\bar{x}^{(1)} = 824.9$ cases per day and $\bar{x}^{(2)} = 818.6$ cases per day.

The value of the test statistic is then calculated to be

$$Z = \frac{824.9 - 818.6}{\sqrt{40/10 + 50/10}} = 2.10.$$

Since $Z > 1.645$ we reject H_0 at a 5% level of significance. The alternative hypothesis H_1 is accepted.

The plant manager concludes that the new production line produces 10 cases per day more than the older line (and may decide to replace more of the older lines as a consequence).

OC Curves for Variances Known

We can also use the OC curves for the normal distribution to find power and sample size for a test. In that case, we use

$$d = \frac{|\mu_1 - \mu_2|}{\sqrt{\sigma_1^2 + \sigma_2^2}}$$

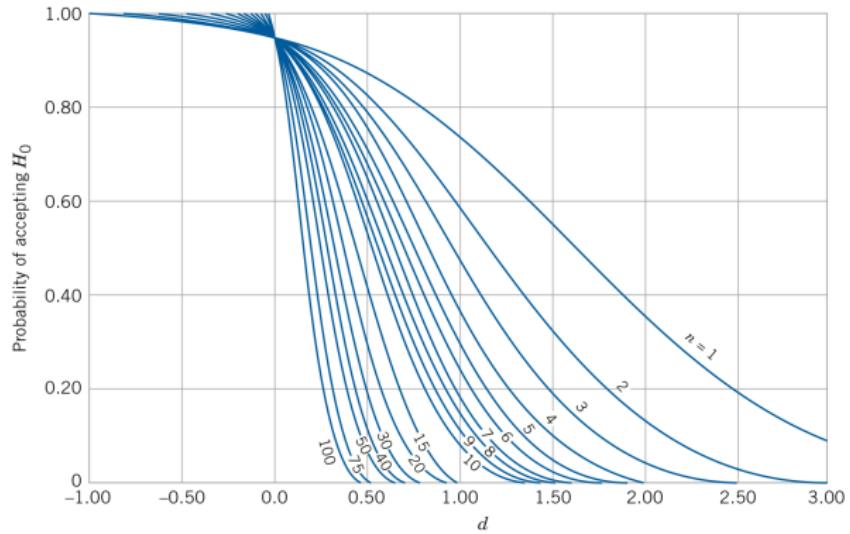
with $n = n_1 = n_2$ (equal sample sizes).

If $n_1 \neq n_2$, the table is used with the **equivalent sample size**

$$n = \frac{\sigma_1^2 + \sigma_2^2}{\sigma_1^2/n_1 + \sigma_2^2/n_2}.$$

OC Curves for Variances Known

21.2. Example. Continuing from Example 21.1, if H_1 is true, we want to find the sample sizes (number of days) required to detect this difference with a probability of 0.90.



We have $d = 10/\sqrt{40 + 50} = 1.05$ and using the chart for $\alpha = 0.05$ (one-sided) we find $n = n_1 = n_2 = 9$.

Confidence Interval for the Difference of Means

21.3. Example. Using the data of Example 21.1, a 95% confidence interval for the difference in mean production is

$$\begin{aligned}\mu_1 - \mu_2 &= \bar{x}^{(1)} - \bar{x}^{(2)} \pm z_{\alpha} \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} \\ &= 824.9 - 818.6 \pm 1.645 \sqrt{40/10 + 50/10} \\ &= 6.3 \pm 4.9\end{aligned}$$

Note that zero is not in this confidence interval, which is expected since $H_0: \mu_1 \leq \mu_2$ was rejected.

Comparing Two Means - Equal Variances

Now suppose that the variances are equal but unknown,

$$\sigma_1^2 = \sigma_2^2 =: \sigma^2.$$

Then

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{\sigma^2(1/n_1 + 1/n_2)}}.$$

is standard normal

Similarly to (19.1), we define the **pooled estimator for the variance**

$$S_p^2 = \frac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}. \quad (21.1)$$

Comparing Two Means - Equal Variances

It is immediately clear that

$$X_{n_1+n_2-2}^2 = \frac{(n_1 + n_2 - 2)S_p^2}{\sigma^2} = \frac{(n_1 - 1)S_1^2}{\sigma^2} + \frac{(n_2 - 1)S_2^2}{\sigma^2}$$

follows a chi-squared distribution with $n_1 + n_2 - 2$ degrees of freedom.

Furthermore,

$$\begin{aligned} T_{n_1+n_2-2} &= \frac{Z}{\sqrt{X_{n_1+n_2-2}^2 / (n_1 + n_2 - 2)}} \\ &= \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)}{\sqrt{S_p^2(1/n_1 + 1/n_2)}} \end{aligned}$$

follows a T -distribution with $n_1 + n_2 - 2$ degrees of freedom.

Confidence Interval for the Difference of Means

We immediately obtain the following $100(1 - \alpha)\%$ confidence interval for $\mu_1 - \mu_2$,

$$(\bar{X}_1 - \bar{X}_2) \pm t_{\alpha/2, n_1+n_2-2} \sqrt{S_p^2(1/n_1 + 1/n_2)},$$

where $t_{\alpha/2, n_1+n_2-2}$ is defined in (13.5).

21.4. Example. In a batch chemical process used for etching circuit boards, two different catalysts are being compared to determine whether they require different emersion times for removal of identical quantities of photo-resistant material.

Twelve batches were run with catalyst 1, resulting in a sample mean emersion time of $\bar{x}_1 = 24.6$ minutes and a sample standard deviation of $s_1 = 0.85$ minutes. Fifteen batches were run with catalyst 2, resulting in a mean emersion time of $\bar{x}_2 = 22.1$ minutes and a standard deviation of $s_2 = 0.98$ minutes.

Confidence Interval for the Difference of Means

We will find a 95% confidence interval on the difference in means $\mu_1 - \mu_2$ assuming that the variances of the two populations are equal. The pooled estimate for the variance gives

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2} = 0.8557$$

so $s_p = 0.925$. Since $t_{0.025, 25} = 2.060$, we obtain

$$\mu_1 - \mu_2 = (2.5 \pm 0.74) \text{ minutes}$$

Student's T -Test for Equal Variances

21.5. Student's T -Test for Equal Variances. Suppose two random samples of sizes n_1 and n_2 from two normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ are given.

Denote by $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ the means of the two samples and let S_p^2 be the pooled sample variance (21.1). Let $(\mu_1 - \mu_2)_0$ be a null value for the difference $\mu_1 - \mu_2$. Then the test based on the statistic

$$T_{n_1+n_2-2} = \frac{(\bar{X}^{(1)} - \bar{X}^{(2)}) - (\mu_1 - \mu_2)_0}{\sqrt{S_p^2(1/n_1 + 1/n_2)}}$$

is called a ***Student's (pooled) test for equality of means.***

We reject at significance level α

- ▶ $H_0: \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ if $|T_{n_1+n_2-2}| > t_{\alpha/2, n_1+n_2-2}$,
- ▶ $H_0: \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0$ if $T_{n_1+n_2-2} > t_{\alpha, n_1+n_2-2}$,
- ▶ $H_0: \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0$ if $T_{n_1+n_2-2} < -t_{\alpha, n_1+n_2-2}$.

Student's *T*-Test for Equal Variances

21.6. Example. Two catalysts are being analyzed to determine how they affect the mean yield of a chemical process. Specifically, catalyst 1 is currently in use, but catalyst 2 is acceptable. Since catalyst 2 is cheaper, if it does not change the process yield significantly, it should be adopted.

We decide to test the hypotheses

$$H_0: \mu_2 \geq \mu_1,$$

$$H_1: \mu_2 \leq \mu_1 - 3\%.$$

From experience with this type of chemical process, the yield follows a normal distribution and the variance of the yield is independent of the catalyst used.

We therefore conduct a Student *T*-test with $\alpha = 5\%$ and choose sample sizes $n_1 = n_2 = 8$. Then the critical value of the test statistic is $t_{0.05,14} = 1.761$.

Student's *T*-Test for Equal Variances

Pilot data yields

$$\bar{x}_1 = 93.75\%, \quad s_1^2 = 3.89\%^2, \quad \bar{x}_2 = 91.73\%, \quad s_2^2 = 4.02\%^2.$$

Then $s_p^2 = 3.96\%^2$ and the test statistic is

$$\frac{\bar{x}_1 - \bar{x}_2}{s_p \sqrt{1/n_1 + 1/n_2}} = 2.03.$$

Since this is greater than the critical value, we reject H_0 and accept H_1 .

We conclude that catalyst 2 induces a significantly lower yield (by 3%) than catalyst 1.

OC Curves for Equal Variances

In the case of equal variances $\sigma_1^2 = \sigma_2^2 = \sigma^2$ and equal sample sizes $n_1 = n_2 = n$, we can use the usual OC curves for the T -test with

$$d = \frac{|\mu_1 - \mu_2|}{2\sigma}.$$

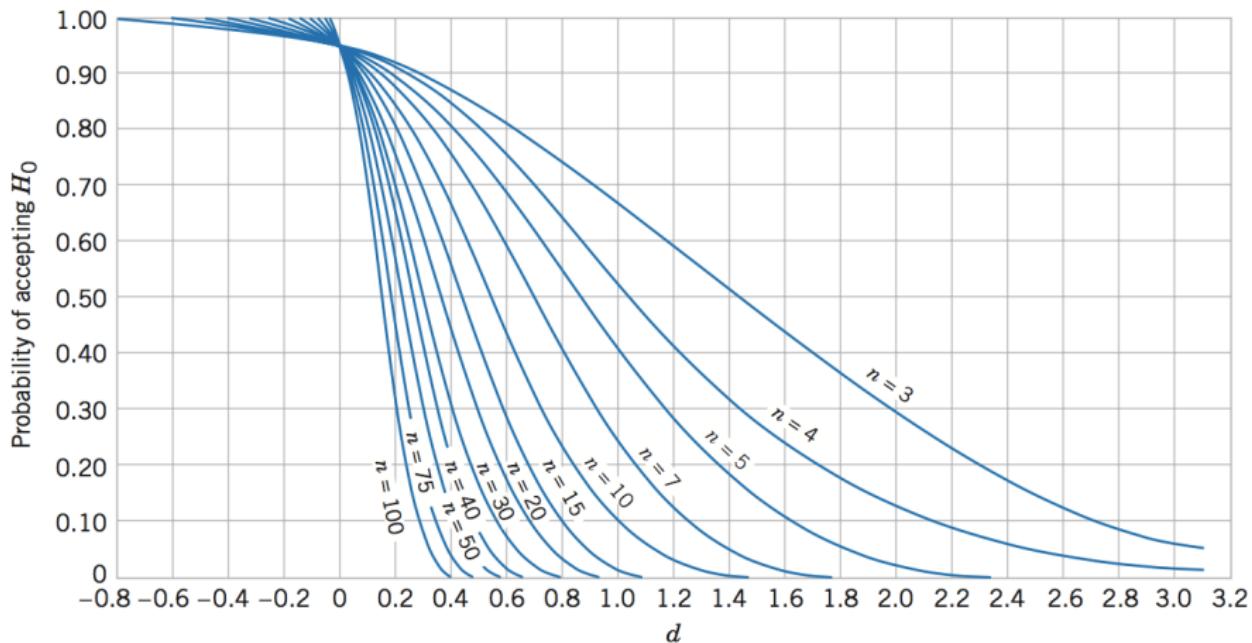
However, we must use the **modified sample size** $n^* = 2n - 1$ when reading the charts. As before, when σ is unknown, we must either use an estimate or express the deviation in terms of σ .

21.7. Example. In setting up the experiment of Example 21.6 it was desired that the power should be at least 0.85. What sample size was required?

We use $\alpha = 0.05$ and the previously determined $s_p = 1.99$ as an estimate for the common standard deviation σ .

OC Curves for Equal Variances

Then $d = \delta/(2\sigma) = 3/(2 \cdot 1.99) = 0.75$ and $\beta = 1 - 0.85 = 0.15$.



The chart gives $n^* = 15$, so $n = (n^* + 1)/2 = 8$ was sufficient.

A Warning Regarding Pre-Testing

The previous discussion of Student's T -test for comparison of means made two assumptions:

- ▶ Both random variables follow normal distributions.
- ▶ Both random variables have equal variances σ^2 .

Comparing the means of two populations is a very common procedure in many applied sciences. In such cases, there is a temptation to

- (i) Collect data.
- (ii) Perform pre-tests on the data (e.g., test for equality of variances or test for normality)
- (iii) Then perform the comparison of means test depending on the result of the pre-test.

This is not recommended!

A Warning Regarding Pre-Testing

Performing such pre-tests and then conditionally on the results using some other test ***on the same data*** will ***invalidate the P-value*** of the comparison of means test.

It is fine to test for normality, equality of variances or other properties and then to ***gather new data for a comparison of means test***. But using the same data creates serious problems.

Literature:

- ▶ Rasch, D., Kubinger, K. and Moder, K. ***The two-sample T test: Pre-testing its assumptions does not pay off.*** Stat. Pap. 52 (2011).
- ▶ Rochon, J., Gondan, M. and Kieser, M. ***To test or not to test: Preliminary assessment of normality when comparing two independent samples.*** BMC Med Res Methodol 12, 81 (2012).
- ▶ Zimmerman, D. W. ***A note on preliminary tests of equality of variances.*** Br J Math Stat Psychol. 57 (2004).

Populations with Unequal Variances

We now consider the case of two normal populations with unequal variances. Recall that

$$\frac{\bar{X}^{(1)} - \bar{X}^{(2)} - (\mu_1 - \mu_2)}{\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}}$$

follows a standard normal distribution.

Now if the variances of the populations are not equal and unknown to us, we are faced with estimating the variance:

$$\widehat{\left(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}\right)} = \frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}.$$

The main problem is that the distribution of the right-hand side is unknown.

The Welch-Satterthwaite Approximation

21.8. Welch-Satterthwaite Relation. Let $X^{(1)}, \dots, X^{(k)}$ be k independent normally distributed random variables with variances $\sigma_1^2, \dots, \sigma_k^2$.

Let s_1^2, \dots, s_k^2 be sample variances based on samples of sizes n_1, \dots, n_k from the k populations, respectively. Let $\lambda_1, \dots, \lambda_k > 0$ be positive real numbers and define

$$\gamma := \frac{(\lambda_1 s_1^2 + \dots + \lambda_k s_k^2)^2}{\sum_{i=1}^k \frac{(\lambda_i s_i^2)^2}{n_i - 1}}.$$

Then

$$\gamma \cdot \frac{\lambda_1 s_1^2 + \lambda_2 s_2^2 + \dots + \lambda_k s_k^2}{\lambda_1 \sigma_1^2 + \lambda_2 \sigma_2^2 + \dots + \lambda_k \sigma_k^2}$$

follows **approximately** a chi-squared distribution with γ degrees of freedom.

The Welch-Satterthwaite Approximation

We are interested in the case $k = 2$, $\lambda_1 = 1/n_1$ and $\lambda_2 = 1/n_2$. Then

$$\gamma = \frac{\left(\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right)^2}{\frac{(S_1^2/n_1)^2}{n_1-1} + \frac{(S_2^2/n_2)^2}{n_2-1}}. \quad (21.2)$$

and

$$\gamma \cdot \frac{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

follows approximately a chi-squared distribution with γ degrees of freedom.
It is then easy to see that

$$T_\gamma = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}}}$$

follows a T -distribution with γ degrees of freedom.

Welch's T -Test for Unequal Variances

21.9. Welch's T -Test for Unequal Variances. Suppose two random samples of sizes n_1 and n_2 from two normal distributions $N(\mu_1, \sigma^2)$ and $N(\mu_2, \sigma^2)$ are given.

Denote by $\bar{X}^{(1)}$ and $\bar{X}^{(2)}$ the means of the two samples and let γ given by (21.2). Let $(\mu_1 - \mu_2)_0$ be a null value for the difference $\mu_1 - \mu_2$. Then the test based on the statistic

$$T_\gamma = \frac{(\bar{X}_1 - \bar{X}_2) - (\mu_1 - \mu_2)_0}{\sqrt{S_1^2/n_1 + S_2^2/n_2}}$$

is called a **Welch's (pooled) test for equality of means**. We reject at significance level α

- ▶ $H_0: \mu_1 - \mu_2 = (\mu_1 - \mu_2)_0$ if $|T_\gamma| > t_{\alpha/2, \gamma}$,
- ▶ $H_0: \mu_1 - \mu_2 \leq (\mu_1 - \mu_2)_0$ if $T_\gamma > t_{\alpha, \gamma}$,
- ▶ $H_0: \mu_1 - \mu_2 \geq (\mu_1 - \mu_2)_0$ if $T_\gamma < -t_{\alpha, \gamma}$.

Welch's T -Test for Unequal Variances

21.10. Remarks.

- ▶ In practice, we **round γ down** to the nearest integer.
- ▶ One disadvantage of unequal variances is that power calculations are much more difficult. There are no simple OC curves for Welch's test.
- ▶ As remarked earlier, it is not a good idea to pre-test for equal variances and then make a decision whether to use Student's or Welch's test. In fact, current recommendations are to **always use Welch's test**. (This is different from what you find in most textbooks. See, for example, the literature below.)

The reason is that Welch's test is only slightly less powerful than Student's test even if the variances are equal. If they are unequal, Student's test is very unreliable.

Literature: The blog article at <http://daniellakens.blogspot.com/2015/01/always-use-welchs-t-test-instead-of.html> and the author's paper cited there.

Non-Parametric Comparisons; Paired Tests and Correlation

Non-Parametric Comparison of Location

Problem: Two independent random variables X and Y are given. Nothing is known about their distribution. Comparing means or even medians is difficult.

Approach: Compare their ***locations*** by checking if

$$P[X > Y] + \frac{1}{2}P[X = Y] \stackrel{?}{=} \frac{1}{2}.$$

If the above probability equals $1/2$, a random observation of X will be greater than or equal to a random observation of Y with probability one-half, and of course the converse is also true.

Here we will assume that X and Y are ***continuous random variables***, so we may omit $P[X = Y]$.

The Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is used to decide whether to reject the null hypothesis

$$H_0: P[X > Y] = \frac{1}{2} \quad \text{or} \quad H_0: P[X > Y] \leq \frac{1}{2}.$$

If both X and Y follow the same distribution, possibly with different location parameter, this may be interpreted as a test comparing the medians of X and Y .

Observations of X and Y are ranked from smallest to largest. For each population, the ranks are summed independently. If $P[X > Y] = 1/2$, then the sum of ranks should be roughly the same for both populations.

To make calculations easier, it is sufficient to consider the sums of ranks of the smaller sample (if sample sizes are different).

The Wilcoxon Rank-Sum Test

22.1. Wilcoxon Rank-Sum Test. Let X and Y be two random samples following some continuous distributions.

Let X_1, \dots, X_m and Y_1, \dots, Y_n , $m \leq n$, be random samples from X and Y and associate the rank R_i , $i = 1, \dots, m + n$, to the R_i th smallest among the $m + n$ total observations. If ties in the rank occur, the mean of the ranks is assigned to all equal values.

Then the test based on the statistic

$$W_m := \text{sum of the ranks of } X_1, \dots, X_m.$$

is called the **Wilcoxon rank-sum test**.

We reject $H_0: P[X > Y] = 1/2$ (and similarly the analogous one-sided hypotheses) at significance level α if W_m falls into the corresponding critical region.

The Wilcoxon Rank-Sum Test

The Wilcoxon rank-sum test is also called the **Mann-Whitney U-test**.
(Often, this refers to the equivalent test where all the ranks, not just those of the smaller sample, are summed.)

For large values of m ($m \geq 20$), W_m is approximately normally distributed with

$$\text{E}[W_m] = \frac{m(m + n + 1)}{2}, \quad \text{Var}[W_m] = \frac{mn(m + n + 1)}{12}.$$

If there are many ties, the variance may be corrected by taking

$$\text{Var}[W_m] = \frac{mn(m + n + 1)}{12 - \sum_{\text{groups}} \frac{t^3 - t}{12}}$$

where the sum is taken over all groups of t ties. However, the best way to deal with ties is still a topic of current research.

Example: Midterm Exam Scores

22.2. Example. It has been suggested that the most highly motivated JI undergraduate students do at least as well, possibly even better, than graduate students in my graduate-level mathematics courses. In the spring term of 2018, there was a significant enrolment of undergraduate students in *Vv557 Methods of Applied Maths II*. The results of the first midterm exam are taken to serve as an indication of the possible truth of this hypothesis.

The hypothesis to be tested is

$$H_0: P[X_{\text{undergrad}} > X_{\text{grad}}] \leq 1/2$$

where $X_{\text{undergrad}}$ and X_{grad} are the exam scores of undergraduate and graduate students enrolled in Vv557.

(The null hypothesis can be interpreted as “undergraduate students do not do better than graduate students in the first midterm.”)

The Raw Data

The following data were recorded (points out of 20 maximum in the first midterm exam):

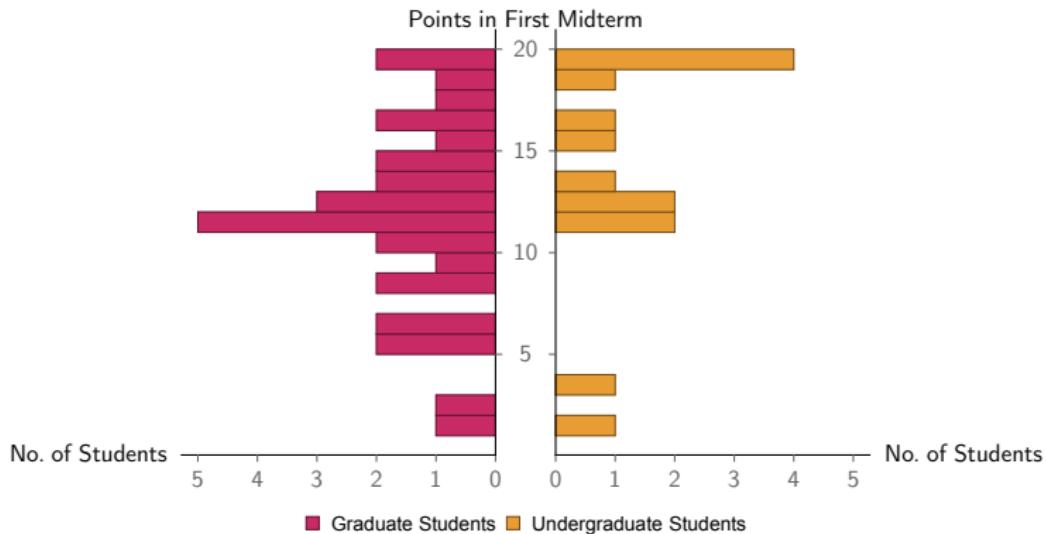
Graduate	5.5	5.5	12.75	18.75	19.25	11.25
	11.5	11.5	12.25	14.25	9.25	14.5
	13.25	8.25	16.75	10.5	6	15.25
	6.5	12.5	10.5	8.75	11.5	17
	2.75	13.25	19	16.5	11.5	1.75
	18.5	12.25	3	15	19.75	11.25
Undergraduate	11.75	19.25	12.25	19.75	16.25	13
	19.25	1.75				

The quartiles are,

Graduate :	8.75,	11.5,	14.5;
Undergraduate :	11.75,	14,	19.25.

Visualizing the Data

The double bar chart below visualizes these data:



Clearly, a non-parametric test is the best choice here.

Ranking the Data

The data are arranged in order and ranked as follows:

Student	Points	Rank	Student	Points	Rank	Student	Points	Rank
grad	1.75	1.5	grad	11.5	17.5	undergrad	15	31
undergrad	1.75	1.5	grad	11.5	17.5	grad	15.25	32
grad	2.75	3	grad	11.5	17.5	undergrad	16.25	33
undergrad	3	4	grad	11.5	17.5	grad	16.5	34
grad	5.5	5.5	undergrad	11.75	20	grad	16.75	35
grad	5.5	5.5	grad	12.25	22	grad	17	36
grad	6	7	undergrad	12.25	22	undergrad	18.5	37
grad	6.5	8	undergrad	12.25	22	grad	18.75	38
grad	8.25	9	grad	12.5	24	grad	19	39
grad	8.75	10	grad	12.75	25	grad	19.25	41
grad	9.25	10	undergrad	13	26	undergrad	19.25	41
grad	10.5	12.5	grad	13.25	27.5	undergrad	19.25	41
grad	10.5	12.5	grad	13.25	27.5	undergrad	19.75	43.5
grad	11.25	14.5	grad	14.25	29	undergrad	19.75	43.5
undergrad	11.25	14.5	grad	14.5	30			

Calculating the Test Statistic

The sum of the ranks of the undergraduate students (smaller sample size) is

$$\begin{aligned}w_{14} &= 1.5 + 4 + 14.5 + 20 + 22 + 22 + 26 \\&\quad + 31 + 33 + 37 + 41 + 41 + 43.5 + 43.5 \\&= 380\end{aligned}$$

Given the large sample sizes, we use a normal approximation for the test statistic (most tables only include values for $m, n \leq 20$). We have

$$E[W_{14}] = \frac{14(14 + 30 + 1)}{2} = 315,$$

$$\text{Var } W_{14} = \frac{14 \cdot 30(14 + 30 + 1)}{12} = 1575$$

Performing the Fisher test

Therefore,

$$Z = \frac{W_m - 315}{\sqrt{1575}}$$

follows a standard normal distribution if $P[X_{\text{undergrad}} > X_{\text{grad}}] = 1/2$. The value of our test statistic is

$$z = \frac{380 - 315}{\sqrt{1575}} = 1.64.$$

Using the normal distribution table, we find a P -value of

$$P[Z \geq 1.64] = 0.0505.$$

There is possibly a small indication that undergraduate students might do better than graduate students, but the evidence is far from conclusive.

Discussion of the Wilcoxon Rank Tests

In the previous example we did not apply the correction for ties to the variance of the normal distribution - why?

Had we done so, the variance would have been ***negative*** - not good!

Could we have used an exact table of critical values? No, because no such table exist for $m, n > 20$. The Wilcoxon rank tests are ***combinatorial tests*** and P -values become increasingly hard to calculate exactly as the number of possible permutations of ranks increases with m and n .

For this reason, ***ties are problematic*** since they increase the number of possible permutations. We have presented one way to deal with ties (assigning the average rank) but this is not the only approach. This is the subject of current research!

Literature: McGee, M. ***Case for omitting tied observations in the two-sample T-test and the Wilcoxon-Mann-Whitney Test.*** PLoS One 13:7, 2018.

Paired Tests

Problem: When comparing means (or, in general, the location) of two populations ***extraneous factors*** may distort the results.

22.3. Example. Suppose we wish to study the efficacy of two different drugs in fighting a disease, Drug A and Drug B. A simple approach would be to treat 20 patients with Drug A and 20 patients with Drug B and then compare the average degree of improvement.

However, it could be that (for example) the disease affects smokers more severely than non-smokers.

If there are more smokers among the sample for Drug A than for Drug B, this could cause the improvements measured for Drug A to be less evident than for Drug B, even if overall Drug A were the better drug.

Paired Tests

Instead, a better approach is to ***pair the samples***: For every person with certain characteristics (gender, age smoker/non-smoker, etc.) administered with Drug A, a person with the same characteristics is administered Drug B.

That means that the sample sizes must be equal in both populations and every sample observation in one population is paired with a corresponding observation in the other population.

Suppose we have two populations with random variables X and Y that we wish to compare. We then define a new random variable

$$D := X - Y$$

and conduct all tests on D .

Paired T -Tests

We note that

$$\mu_D = E[D] = E[X - Y] = E[X] - E[Y] = \mu_X - \mu_Y.$$

Therefore, the hypothesis (for example)

$$H_0: \mu_X = \mu_Y \quad \text{may be replaced with} \quad H_0: \mu_D = 0.$$

We will assume that X and Y follow a **joint bivariate normal distribution**. Then it is not hard to see that $D = X - Y$ follows a normal distribution.

We then consider a paired random sample $(X_1, Y_1), \dots, (X_n, Y_n)$ from both populations yielding a sample D_1, \dots, D_n with $D_i := X_i - Y_i$, $i = 1, \dots, n$.

We denote by \bar{D} the sample mean and by S_D^2 the sample variance of D .

Paired T -Tests

Then

$$T_{n-1} = \frac{\bar{D} - \mu_D}{\sqrt{S_D^2/n}}$$

follows a T -distribution with $n - 1$ degrees of freedom.

We may find confidence intervals for μ_D and conduct hypothesis tests as we would for any normally distributed random variable. A T -test for D is called a ***paired T-test*** for Y and Y' .

Paired T -Tests

22.4. Example. In a study of the effectiveness of physical exercise in weight reduction, a group of 16 persons engaged in a prescribed program of physical exercise for one month showed the following results :

Weight before (X)	209	178	169	212	180	192	158	180
Weight after (Y)	196	171	170	207	177	190	159	180
<hr/>								
$D = Y - X$	-13	-7	+1	-5	-3	-2	+1	0
<hr/>								
Weight before (X)	170	153	183	165	201	179	243	144
Weight after (Y)	164	152	179	162	199	173	231	140
<hr/>								
$D = Y - X$	-6	-1	-4	-3	-2	-6	-12	-4

We want to test at the 0.01 level of significance whether the exercise program is effective.

Paired T -Tests

We decide to test

$$H_0: \mu_D \geq 0.$$

From the $n = 16$ data we have $\bar{D} = -4.125$ and $s_D^2 = 16.517$. The test statistic is

$$T = \frac{\bar{D}}{s_D / \sqrt{n}} = -4.06.$$

Since $t_{0.01,15} = 2.602$, we may reject H_0 at the 0.01 level of significance.

There is evidence that the physical exercise program leads to a loss of weight.

Non-Parametric Paired Test

Suppose the two independent random variables X and Y do not follow a normal distribution. Then we would like to treat $D = X - Y$ by the Wilcoxon signed-rank test.

The signed-rank test requires a random variable to have a **symmetric** distribution. What does that mean?

A random variable X is said to be **symmetric about $a \in \mathbb{R}$** if

$$X - a \quad \text{and} \quad -(X - a)$$

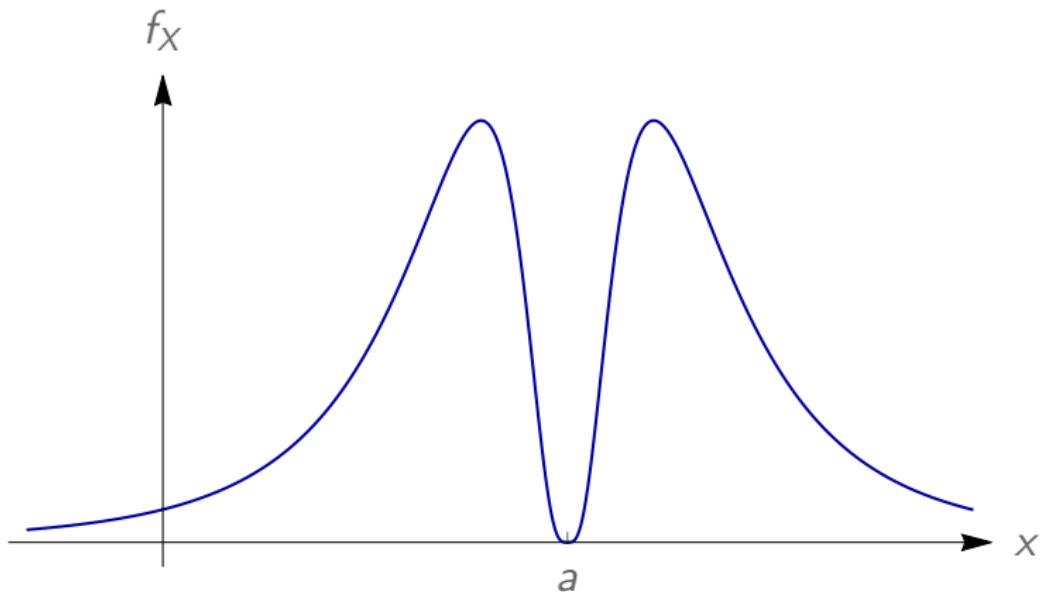
have the same distribution.

In terms of the density function f_X this means that

$$f_X(x - a) = f_X(a - x)$$

(as can be verified by applying Theorem 7.5.)

Example of the Density of a Symmetric Distribution



Properties of $D = X - Y$

Now let X and Y be two independent random variables that follow the same distribution but differ only in their location, i.e., $X' := X - \delta$ and Y are independent and identically distributed.

Then

$$P[X - Y > \delta] = P[X - \delta - Y > 0] = P[X' - Y > 0] = \frac{1}{2}$$

so δ is the median of $X - Y$.

Furthermore

$$D = X - Y = \delta + X' - Y$$

and

$$2\delta - D = \delta + Y - X'$$

have the same distribution since Y and X' are i.i.d. random variables.

Non-Parametric Paired Test

Therefore, D will be symmetric about its median δ and we can apply the Wilcoxon signed rank test to test hypotheses about δ .

Historically, Wilcoxon proposed both the rank-sum test (for pooled comparisons) and the signed-rank test (for paired comparisons) in a single publication.

Paired vs. Pooled T -Tests

Let us take another look at the test statistics for a paired T -test. We note that

$$\begin{aligned}\bar{D} &= \frac{1}{n} \sum_{i=1}^n (X_i - Y_i) \\ &= \frac{1}{n} \sum_{i=1}^n X_i - \frac{1}{n} \sum_{i=1}^n Y_i \\ &= \bar{X} - \bar{Y}\end{aligned}$$

and $\mu_D = \mu_X - \mu_Y$. This suggests that the paired and pooled T -test may actually be fairly similar.

Paired vs. Pooled T -Tests

For our comparison, let us assume that we have two populations of normally distributed random variables X and Y with equal variances σ^2 .

We want to test

$$H_0: \mu_X = \mu_Y,$$

and take a paired sample of equal size n from (X, Y) .

Then we could either perform a paired test or a pooled test - which is more powerful? Let us compare the test statistics:

$$T_{\text{pooled}} = \frac{\bar{X} - \bar{Y}}{\sqrt{2S_p^2/n}}, \quad \text{critical value} = t_{\alpha/2, 2n-2},$$

$$T_{\text{paired}} = \frac{\bar{X} - \bar{Y}}{\sqrt{S_D^2/n}}, \quad \text{critical value} = t_{\alpha/2, n-1},$$

Paired vs. Pooled T -Tests

We immediately note that

the pooled test has more degrees of freedom

and so rejecting H_0 is easier - the test would be more powerful, if the test statistics were equal.

But the test statistics differ:

- ▶ In the pooled test, the denominator contains

$$2S_p^2/n \quad \text{which estimates} \quad 2\sigma^2/n.$$

- ▶ In the pooled test, the denominator contains

$$S_D^2/n \quad \text{which estimates} \quad \sigma_D^2/n = \sigma_{\bar{D}}^2.$$

Paired vs. Pooled T -Tests

To discuss the two denominators, we will compare

$$\frac{2\sigma^2}{n} \quad \text{with} \quad \sigma_{\bar{D}}^2.$$

A direct calculation yields

$$\begin{aligned}\sigma_{\bar{D}}^2 &= \text{Var}[\bar{D}] \\ &= \text{Var}[\bar{X} - \bar{Y}] \\ &= \text{Var}[\bar{X}] + \text{Var}[\bar{Y}] - 2 \text{Cov}[\bar{X}, \bar{Y}] \\ &= \frac{\sigma^2}{n} + \frac{\sigma^2}{n} - 2 \frac{\sigma^2}{n} \frac{\text{Cov}[\bar{X}, \bar{Y}]}{\sqrt{\text{Var}[\bar{X}]}\sqrt{\text{Var}[\bar{Y}]}} \\ &= \frac{2\sigma^2}{n}(1 - \rho_{\bar{X}\bar{Y}})\end{aligned}$$

where $\rho_{\bar{X}\bar{Y}}$ is the correlation coefficient of X and Y .

Correlation Coefficient of Sample Means

It is worth doing a quick calculation to verify that in our case (paired samples)

$$\rho_{\bar{X} \bar{Y}} = \rho_{XY}.$$

Since the covariance is bilinear,

$$\begin{aligned}\text{Cov}[\bar{X}, \bar{Y}] &= \text{Cov}\left[\frac{1}{n} \sum_{i=1}^n X_i, \frac{1}{n} \sum_{j=1}^n Y_j\right] \\ &= \frac{1}{n^2} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}[X_i, Y_j] \\ &= \frac{1}{n^2} \sum_{i=1}^n \text{Cov}[X_i, Y_i]\end{aligned}$$

where we have used that X_i and Y_j are independent for $i \neq j$.

Correlation Coefficient of Sample Means

Then $\text{Cov}[X_i, Y_i] = \text{Cov}[X, Y]$, so

$$\text{Cov}[\bar{X}, \bar{Y}] = \frac{1}{n} \text{Cov}[X, Y].$$

and, therefore,

$$\begin{aligned}\rho_{\bar{X} \bar{Y}} &= \frac{\text{Cov}[\bar{X}, \bar{Y}]}{\sqrt{\text{Var}[\bar{X}]}\sqrt{\text{Var}[\bar{Y}]}} \\ &= \frac{\frac{1}{n} \text{Cov}[X, Y]}{\sqrt{\text{Var}[X]/n}\sqrt{\text{Var}[Y]/n}} \\ &= \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X]}\sqrt{\text{Var}[Y]}} \\ &= \rho_{XY}.\end{aligned}$$

Paired vs. Pooled T -Tests

The upshot of all this is that

$$\sigma_D^2 = \frac{2\sigma^2}{n}(1 - \rho_{XY}).$$

Therefore, if $\rho_{XY} > 0$, the denominator of the paired statistic will be smaller than that of the pooled statistic, leading to a larger value of the statistic and a higher power of the test.

On the other hand, if ρ_{XY} is zero (or even negative), then pairing is intuitively unnecessary and in fact causes the test to lose power, since it is easier to reject H_0 when comparing with $t_{\alpha/2, 2n-2}$ than with $t_{\alpha/2, n-1}$.

Pairing in the absence of correlation makes a test less powerful.

Estimating Correlation

Since correlation is important in deciding whether to use a paired or a pooled T -test, let us briefly discuss the estimation of ρ .

Let us take a random sample of size n from (X, Y) as before. Then we have the natural unbiased estimators

$$\widehat{\text{Var}[X]} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2,$$

$$\widehat{\text{Cov}[X, Y]} = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})$$

The natural choice (method of moments!) for an estimator for the correlation coefficient is then

$$R := \hat{\rho} = \frac{\sum (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum (X_i - \bar{X})^2} \sqrt{\sum (Y_i - \bar{Y})^2}}. \quad (22.1)$$

Correlation of Bivariate Normal Random Variables

Now let us suppose that (X, Y) follows a bivariate normal distribution, i.e., they have the joint density

$$f_{XY}(x, y) = \frac{1}{2\pi\sigma_X\sigma_Y\sqrt{1-\varrho^2}} e^{-\frac{1}{2(1-\varrho^2)} \left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\varrho\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2 \right]}$$

with $\mu_X, \mu_Y \in \mathbb{R}$, $\sigma_X, \sigma_Y > 0$ and correlation coefficient $\varrho \in (-1, 1)$.

Under this assumption, we will introduce a hypothesis test and a confidence interval for the correlation coefficient.

An important role is played by the Fisher transformation (8.3).

Hypothesis Tests for the Correlation Coefficient

It can be shown that for large n the Fisher transformation of R ,

$$\frac{1}{2} \ln \left(\frac{1+R}{1-R} \right) = \text{Artanh}(R)$$

is approximately normally distributed with

$$\mu = \frac{1}{2} \ln \left(\frac{1+\varrho}{1-\varrho} \right) = \text{Artanh}(\varrho), \quad \sigma^2 = \frac{1}{n-3}.$$

We can thus test $H_0: \varrho = \varrho_0$, by using the test statistic

$$\begin{aligned} Z &= \frac{\sqrt{n-3}}{2} \left(\ln \left(\frac{1+R}{1-R} \right) - \ln \left(\frac{1+\varrho_0}{1-\varrho_0} \right) \right) \\ &= \sqrt{n-3} (\text{Artanh}(R) - \text{Artanh}(\varrho_0)) \end{aligned} \tag{22.2}$$

Confidence Interval for the Correlation Coefficient

Furthermore, from (22.2) we can calculate a $100(1 - \alpha)\%$ confidence interval for ϱ , given explicitly by

$$\left[\frac{1 + R - (1 - R)e^{2z_{\alpha/2}/\sqrt{n-3}}}{1 + R + (1 - R)e^{2z_{\alpha/2}/\sqrt{n-3}}}, \frac{1 + R - (1 - R)e^{-2z_{\alpha/2}/\sqrt{n-3}}}{1 + R + (1 - R)e^{-2z_{\alpha/2}/\sqrt{n-3}}} \right].$$

or

$$\tanh \left(\text{Artanh}(R) \pm \frac{z_{\alpha/2}}{\sqrt{n-3}} \right).$$

Correlation as a Measure of Skill vs. Luck

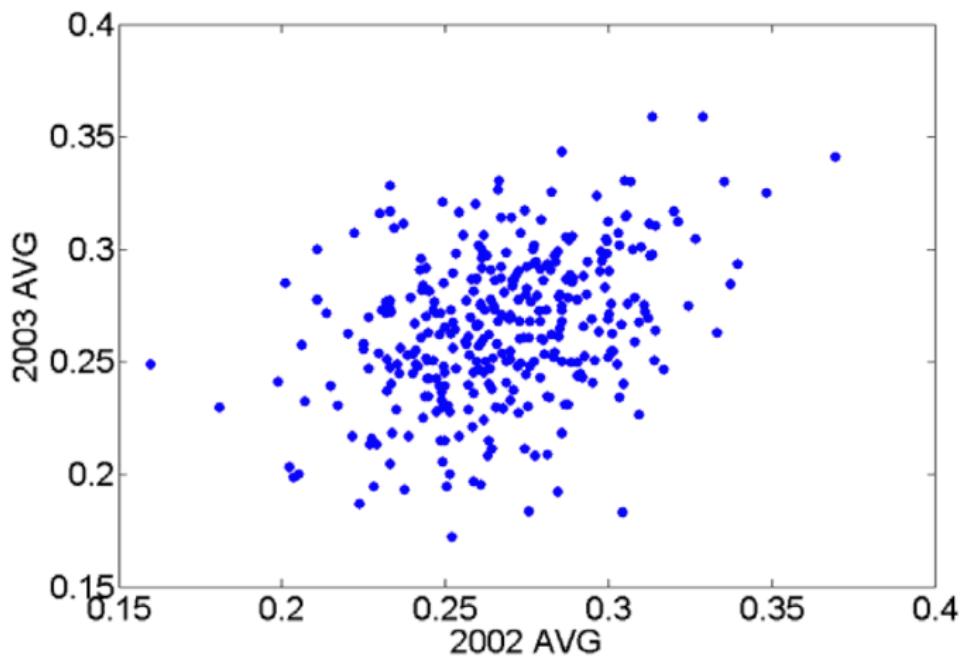
22.5. Example. The article ***A Batting Average: Does It Represent Ability or Luck?*** explores the suitability of a baseball player's "batting average" (BA for short; the number of hits divided by the number of at-bats) as a measure of skill.

The premise is the following:

- ▶ If a good BA is a matter of skill, a player's BA will have a consistent value from one year to the next. We would expect the batting average of a random player in one year to be linearly correlated to the BA in the next year.
- ▶ If a good BA is a matter of luck, a player's BA will vary from one year to the next and the BA as a random variable in a given year will be uncorrelated or even independent of the BA in another year.

We show the batting averages of all players with at least 100 at-bats in the 2002 and 2003 seasons on the next slide.

Correlation as a Measure of Skill vs. Luck

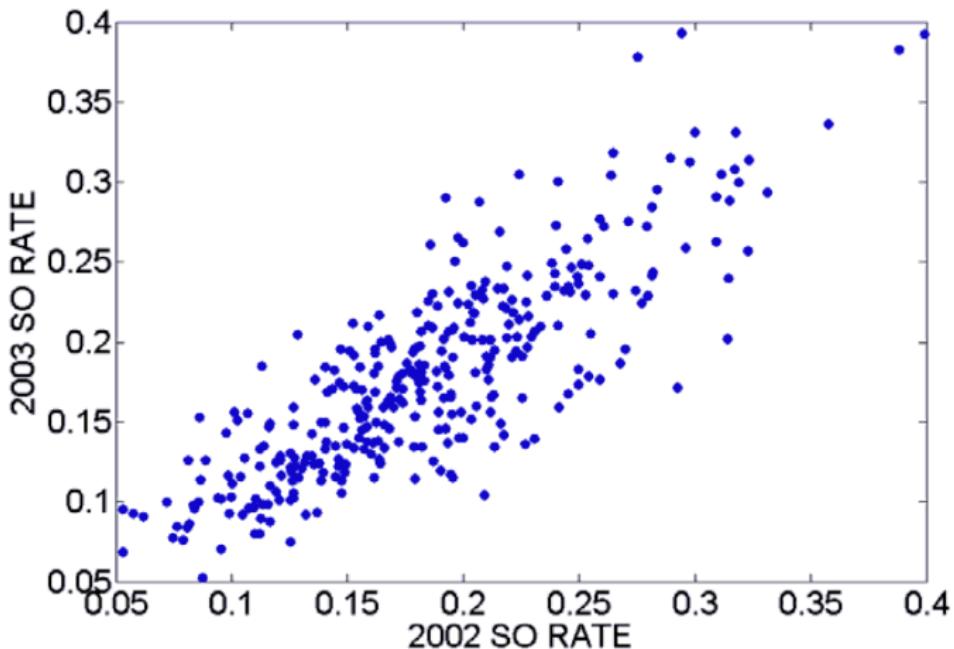


Batting averages in the 2002 and 2003 baseball seasons. J. Albert. *A Batting Average: Does It Represent Ability or Luck?*

The article proposes that the “strikeout rate” is a better measure of skill.

Correlation as a Measure of Skill vs. Luck

Indeed, the corresponding scattergram (for the same players) seems to exhibit a stronger linear dependence from one year to the next:



Categorical Data

Categorical Data

Problem: Instead of assuming numerical values, data may fall into categories. Such data is called **categorical data**.

23.1. Example. Mars Corporation's **M&Ms** are produced in different colors: red, green, blue, brown, yellow and orange. If we pick a random M&M, it will randomly have one of these colors.

Approach: Each member of a population falls into one of k given categories with probability p_i , $0 < p_i < 1$, $i = 1, \dots, n$, and

$$p_1 + p_2 + \cdots + p_k = 1.$$

Our goal is to make inferences on the values of these p_i .

Categorical Random Variables

We suppose that a random variable X is given, where X can take on the values $1, \dots, k$ with respective probabilities p_1, \dots, p_k as above. We say that X is a **categorical random variable**.

A random sample of size n from X is collected and the results are expressed as a random vector

$$(X_1, X_2, \dots, X_k) \quad \text{with} \quad X_1 + X_2 + \cdots + X_k = n.$$

For example, a packet containing $n = 14$ M&M's will yield a random vector $(X_{\text{red}}, X_{\text{green}}, \dots, X_{\text{orange}})$.

When $k = 2$, then the distribution governing the probability of an item falling into category 1 ("success") or category 2 ("failure") is the binomial distribution. For $k > 2$, we need to develop a new distribution.

Multinomial Trials

23.2. Definition. A ***multinomial trial*** with parameters p_1, \dots, p_k is a trial that can result in exactly one of k possible outcomes. The probability that outcome i will occur on a given trial is p_i , for $i = 1, \dots, k$.

23.3. Remark. It is clear from the definition that $0 \leq p_i \leq 1$, $i = 1, \dots, k$, and $p_1 + \dots + p_k = 1$. To avoid unnecessary trivial cases, we assume $0 < p_i < 1$, $i = 1, \dots, k$.

For $k = 2$, $p_1 = p$ and $p_2 = q = 1 - p$, we regain the classic Bernoulli trial.

A ***multinomial random variable*** now counts the number of times that outcome i occurs when a fixed number of n i.i.d. multinomial trials is performed. It therefore generalizes the binomial random variable.

The Multinomial Distribution

23.4. Definition. A random vector $((X_1, \dots, X_k), f_{X_1 X_2 \dots X_k})$ with

$$(X_1, \dots, X_k) : S \rightarrow \Omega = \{0, 1, 2, \dots, n\}^k$$

and (joint) distribution function $f_{X_1 X_2 \dots X_k} : \Omega \rightarrow \mathbb{R}$ given by

$$f_{X_1 X_2 \dots X_k}(x_1, \dots, x_k) = \frac{n!}{x_1! \dots x_k!} p_1^{x_1} \cdots p_k^{x_k},$$

$p_1, \dots, p_k \in (0, 1)$, $n \in \mathbb{N} \setminus \{0\}$ is said to have a **multinomial distribution** with parameters n and p_1, \dots, p_k .

23.5. Remark. Of course, it would be sufficient to consider a $k - 1$ dimensional random variable, as one of the X_i is wholly determined by the others. (The case $k = 3$ is handled by a bivariate, the case $k = 2$ by a simple random variable.) For reasons of symmetry it is, however, worth investing in the additional random variable.

Expectation and Variance of the Multinomial Distribution

23.6. Theorem. Let $((X_1, \dots, X_k), f_{X_1, X_2, \dots, X_k})$ be a multinomial random variable with parameters n and p_1, \dots, p_k .

- (i) The (marginal) expectations of the individual random variables X_i are given by

$$\mathbb{E}[X_i] = np_i, \quad i = 1, \dots, k.$$

- (ii) $\text{Var}[X_i] = np_i(1 - p_i)$, $i = 1, \dots, k$,
(iii) $\text{Cov}[X_i, X_j] = -np_i p_j$, $1 \leq i < j \leq k$.

While results (i) and (ii) are easy to see, the proof of (iii) requires some work. Since we won't need that result, it is left to you.

The Pearson Statistic

Hypothesis testing and statistical analysis are based on the following result, which we will not prove:

23.7. Theorem. Let $((X_1, \dots, X_k), f_{X_1 X_2 \dots X_k})$ be a multinomial random variable with parameters n and p_1, \dots, p_k . For large n the **Pearson statistic**

$$\sum_{i=1}^k \frac{(X_i - np_i)^2}{np_i} \tag{23.1}$$

follows an approximate chi-squared distribution with $k - 1$ degrees of freedom.

The Pearson Statistic

A good way to memorize this statistic is to see that the X_i are the “observed” frequencies and $np_i = E[X_i]$ are the “expected” frequencies.

Writing $O_i := X_i$ and $E_i := E[X_i]$, (23.1) becomes

$$\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

23.8. Remark. The number of degrees of freedom in Theorem 23.7 is equal to the number of independent cells: given k cells and a total of n multinomial trials, the number of results in the first $k - 1$ cells is random, while the number of results in the final cell is completely determined by these $k - 1$ results.

One could say that there are $k - 1$ independent cells, hence $k - 1$ degrees of freedom.

Cochran's Rule

In the context of Theorem 23.7 we need to know how large n needs to be for the chi-squared distribution to be a good approximation to the true distribution of the statistic (23.1).

Cochran's Rule states that we should require

$$\mathbb{E}[X_i] = np_i \geq 1 \quad \text{for all } i = 1, \dots, k,$$

$$\mathbb{E}[X_i] = np_i \geq 5 \quad \text{for 80\% of all } i = 1, \dots, k,$$

Especially if the p_i are not known roughly beforehand, care needs to be taken to ensure that the sample size n is sufficiently large so that these criteria can apply.

Literature: Kroonenberg, P. M. and Verbeek, A. *The Tale of Cochran's Rule: My Contingency Table has so Many Expected Values Smaller than 5, What Am I to Do?*, The American Statistician, 72:2 (2018)



William G. Cochran (1909-1980)
Statisticians in History, Amstat News
(2016)

Fisher Test for Multinomial Distribution

We can then develop a statistical test for the hypothesis that a set of data follows a given multinomial distribution: if data follows a given distribution, the number of observed values in each category will be close to the expected number and (23.1) will be small. Conversely, if (23.1) is large and the observed data deviates from the expected data significantly, then we have evidence that the data does not follow the presumed distribution.

This test will, by its nature, always be a Fisher test. Furthermore, it makes no sense to use terms such as “two-sided” or “ones-sided” for the test, since if some p_i are larger than their null values, then some other p_i will be smaller than their null values.

Test for Multinomial Distribution

23.9. Pearson's Chi-squared Goodness-of-Fit Test. Let (X_1, \dots, X_k) be a sample of size n from a categorical random variable with parameters (p_1, \dots, p_k) satisfying Cochran's Rule. Let $(p_{i_0}, \dots, p_{k_0})$ be a vector of null values. Then the test

$$H_0: p_i = p_{i_0}, \quad i = 1, \dots, k,$$

based on the statistic

$$\chi^2_{k-1} = \sum_{i=1}^k \frac{(X_i - np_{i_0})^2}{np_{i_0}}$$

is called an **chi-squared goodness-of-fit test**.

We reject H_0 at significance level α if $\chi^2_{k-1} > \chi^2_{\alpha, k-1}$.

Multinomial Statistics

23.10. Example. A computer scientist has developed an algorithm for generating pseudorandom integers over the interval 0-9. He codes the algorithm and generates 1000 pseudorandom digits. The data is shown below:

i	0	1	2	3	4	5	6	7	8	9	n
O_i	94	93	112	101	104	95	100	99	94	108	1000
E_i	100	100	100	100	100	100	100	100	100	100	1000

We want to test whether these data conform to a discrete uniform distribution on $\Omega = \{0, 1, 2, \dots, 9\}$.

Formally, we test

H_0 : The data follow a multinomial distribution

with parameters $(p_0, \dots, p_9) = \left(\frac{1}{10}, \dots, \frac{1}{10}\right)$.

Multinomial Statistics

The observed test statistic is

$$\sum_{i=0}^9 \frac{(O_i - E_i)^2}{E_i} = \frac{(94 - 100)^2}{100} + \cdots + \frac{(108 - 100)^2}{100} = 3.72$$

This statistic follows a chi-squared distribution with $10 - 1 = 9$ degrees of freedom. Since $\chi^2_{0.05,9} = 16.92$, the P -value of the test is greater than 5%. We decide not to reject H_0 .

We conclude that there is no evidence that the generated numbers are not random.

Goodness-of-Fit Test for a Discrete Distribution

The previous discussion centers on testing whether categorical data conforms to a ***completely determined*** distribution, i.e., we compare directly to a multinomial distribution. However, we can also use this method to see whether data conforms to an arbitrary discrete or continuous distribution.

Such a distribution will typically have one or more parameters, which we also estimate from the given data. We also need to divide our data into categories, so we can use our multinomial test. If we use our data to estimate parameters of the distribution, the statistic

$$\sum_{i=1}^k \frac{(O_i - E_i)^2}{E_i}$$

will follow a chi-squared distribution with $k - 1 - m$ degrees of freedom, where m is the number of parameters that we estimate.

Goodness-of-Fit Test for a Discrete Distribution

23.11. Example. It is claimed that the number of defects in printed circuit boards follows a Poisson distribution with unknown parameter k . We want to determine if there is evidence that this claim is false.

A random sample of $n = 60$ printed boards has been collected and the following number of defects observed:

Number of Defects X	Observed Frequency
0	32
1	15
2	9
3	4

The parameter k (which is also the mean) of the assumed Poisson distribution is unknown and must be estimated from the data.

Goodness-of-Fit Test for a Discrete Distribution

From Example 12.5 we know that a maximum-likelihood estimator for k is the sample mean,

$$\hat{k} = \bar{X} = \frac{1}{60}(32 \cdot 0 + 15 \cdot 1 + 9 \cdot 2 + 4 \cdot 3) = 0.75.$$

In order to apply the multinomial distribution, we first calculate

$$P[X = 0] = \frac{e^{-\hat{k}} \hat{k}^0}{0!} = 0.472$$

$$P[X = 1] = \frac{e^{-\hat{k}} \hat{k}^1}{1!} = 0.354$$

$$P[X = 2] = \frac{e^{-\hat{k}} \hat{k}^2}{2!} = 0.133$$

$$P[X \geq 3] = 1 - P[X = 0] - P[X = 1] - P[X = 2] = 0.041$$

Goodness-of-Fit Test for a Discrete Distribution

We can therefore replace the distribution of X with that of a categorical random variable with parameters

$$(p_0, p_1, p_2, p_3) = (0.472, 0.354, 0.133, 0.041).$$

We calculate the expected frequencies $E_i = np_i$ as follows:

Number of Defects X (Category i)	Expected Frequency E_i
0	$60 \cdot 0.472 = 28.32$
1	$60 \cdot 0.354 = 21.24$
2	$60 \cdot 0.133 = 7.98$
3	$60 \cdot 0.041 = 2.46$

We see that $E_3 < 5$ and since we have only four categories, this means that more than 1 in 5 categories have an expected frequency smaller than 5. Since Cochran's Rule is not satisfied, we may not apply Pearson's test.

Goodness-of-Fit Test for a Discrete Distribution

The problem can be solved by combining the last two categories:

Category i	Exp. Frequency E_i	Obs. Frequency O_i
0	28.32	32
1	21.24	15
2	10.44	13

The test

H_0 : the number of defects follows a Poisson distribution with parameter $k = 0.75$

is then equivalent to the test

H_0 : the number of defects follows a multinomial distribution with parameters $(0.472, 0.354, 0.174)$

Goodness-of-Fit Test for a Discrete Distribution

For $N = 3$ categories, the statistic

$$\chi^2 = \sum_{i=1}^N \frac{(O_i - E_i)^2}{E_i}$$

then follows a chi-squared distribution with $N - 1 - m = 3 - 1 - 1 = 1$ degree of freedom. Now

$$\chi^2 = \frac{(32 - 28.32)^2}{28.32} + \frac{(15 - 21.24)^2}{21.24} + \frac{(13 - 10.44)^2}{10.44} = 2.94,$$

and the critical value for $\alpha = 0.05$ is $\chi^2_{0.05,1} = 3.84$. Since the observation does not lie in the critical region, we are unable to reject H_0 at the 5% level of significance.

We can also test whether data fits a continuous distribution. In that case, the division of the data range into categories is essentially arbitrary, as illustrated in the following example.

Goodness-of-Fit Test for a Continuous Distribution

23.12. Example. A manufacturing engineer is testing a power supply used in a word processing work station. He wishes to determine whether output voltage is adequately described by a normal distribution. From a random sample of $n = 100$ units he obtains sample estimates of the mean and standard deviation $\bar{x} = 12.04 \text{ V}$ and $s = 0.08 \text{ V}$.

A common practice in constructing the class frequency distribution used in the chi-squared goodness-of-fit test is to choose the category boundaries so that the expected frequencies $E_i = np_i$ are equal for all categories. To use this method, we want to choose the category boundaries a_0, \dots, a_k for the k categories so that all the probabilities

$$p_i = P[a_{i-1} \leq X \leq a_i] = \int_{a_{i-1}}^{a_i} f(x) dx$$

are equal.

Goodness-of-Fit Test for a Continuous Distribution

Suppose we decide to use $k = 8$ cells. For the standard normal distribution the intervals that divide the scale into 8 equally likely segments are

$$\begin{aligned}(a_0, a_1) &= (-\infty, -1.15), & [a_1, a_2) &= [-1.15, -0.675), \\[a_2, a_3) &= [-0.675, -0.32), & [a_3, a_4) &= [-0.32, 0) \\[a_4, a_5) &= [0, 0.32), & [a_5, a_6) &= [0.32, 0.675), \\[a_6, a_7) &= [0.675, 1.15), & [a_7, a_8) &= [1.15, \infty)\end{aligned}$$

For the problem at hand, we need to transform these intervals to corresponding intervals for a normal distribution with mean \bar{x} and standard deviation s . This is easily done by setting

$$a'_i := \bar{x} + sa_i, \quad i = 0, \dots, 8.$$

For each interval, $p_i = 1/8$ so the expected cell frequencies are $E_i = np_i = 100/8 = 12.5$.

Goodness-of-Fit Test for a Continuous Distribution

The engineer observes 100 voltages as given below:

Category i	Exp. Frequency E_i	Obs. Frequency O_i
$x < 11.948$	12.5	10
$11.948 \leq x < 11.986$	12.5	14
$11.986 \leq x < 12.014$	12.5	12
$12.014 \leq x < 12.040$	12.5	13
$12.040 \leq x < 12.066$	12.5	11
$12.066 \leq x < 12.094$	12.5	12
$12.094 \leq x < 12.132$	12.5	14
$12.132 \leq x$	12.5	14
	100	100

We calculate

$$\chi^2 = \sum_{i=1}^8 \frac{(O_i - E_i)^2}{E_i} = 1.12$$

Goodness-of-Fit Test for a Continuous Distribution

We have $k = 8$ categories. Since two parameters in the normal distribution have been estimated, this statistic follows a chi-squared distribution with $k - 1 - m = 8 - 1 - 2 = 5$ degrees of freedom.

We see find that $\chi^2_{0.95,5} = 1.15$, so the P -value of the test is greater than 95%.

We conclude that there is no reason to believe that output voltage is not normally distributed.

Goodness-of-Fit Tests with Mathematica

The goodness-of-fit test is also implemented in Mathematica. However, there is no control over the number of categories chosen; Mathematica will always choose about $2n^{2/5}$ categories and ignore the condition (??). Hence the test will not always be reliable. For instance, using the data from Example 23.11, we have

```
Needs["HypothesisTesting`"];
data := Join[Table[0, {i, 1, 32}], Table[1, {i, 1, 15}],
    Table[2, {i, 1, 9}], Table[3, {i, 1, 4}]];
PearsonChiSquareTest[data, PoissonDistribution[k],
 {"FittedDistributionParameters", "DegreesOfFreedom",
 "TestDataTable"}]
```

$$\left\{ \{k \rightarrow 0.75\}, 9, \frac{\begin{array}{c|cc} \text{Statistic} & \text{P-Value} \\ \hline \end{array}}{\text{Pearson } \chi^2} \right\} \quad \left\{ \begin{array}{cc} 3.46021 & 0.177266 \end{array} \right\}$$

Mathematica uses $k = \lceil 2 \cdot 60^{2/5} \rceil = 11$ categories even though the data only occurs in four categories and Cochran's Rule isn't satisfied. The test statistic is the same as ours would have been had we used all four categories.

Independence of Categorizations

The multinomial distribution and the Pearson statistic are also very useful in another situation, best explained by an example:

23.13. Example. A researcher wants to study the relationship between nightly hours of sleep and academic performance of students. A test group of students fills out a questionnaire, giving their amount of sleep and their current GPA score. The test group can then be divided as follows

$$\{\text{test group}\} = \{\text{< 6h sleep}\} \cup \{\text{6-9h sleep}\} \cup \{\text{> 9h sleep}\},$$

$$\{\text{test group}\} = \{\text{low GPA}\} \cup \{\text{average GPA}\} \cup \{\text{high GPA}\}$$

If academic performance and nightly sleep are not related to each other, these categorizations will be independent, i.e., the likelihood of a student falling into any of the GPA categories will not depend on which sleep category the student is in.

Contingency Tables

The data from the test group can be summarized in a **contingency table** as follows:

	< 6h sleep	6-9h sleep	> 9h sleep
low GPA	n_{11}	n_{12}	n_{13}
average GPA	n_{21}	n_{22}	n_{23}
high GPA	n_{31}	n_{32}	n_{33}

Every member of the test group will count as 1 member of a specific **cell** (table entry) and the cells list the number of members with the corresponding properties. For example,

n_{23} = number of students with average GPA

and more than 9 hours of nightly sleep.

$r \times c$ Contingency Tables; Marginal Sums

In general, we will treat situations where the contingency table has r rows and c columns. We define the **marginal row and column sums**

$$n_{i\cdot} = \sum_{j=1}^c n_{ij},$$

$$n_{\cdot j} = \sum_{i=1}^r n_{ij}.$$

In our example,

	< 6h sleep	6-9h sleep	> 9h sleep	
low GPA	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$
average GPA	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$
high GPA	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n

Cell Probabilities and Independence

Suppose that

- ▶ p_{ij} is the probability of falling into the cell of the i th row and the j th column,
- ▶ $p_{i\cdot}$ is the probability of falling anywhere in the i th row,
- ▶ $p_{\cdot j}$ is the probability of falling anywhere in the j th column.

If the row and column categorizations are independent, then it should be the case that

$$H_0: p_{ij} = p_{i\cdot} \cdot p_{\cdot j}. \quad (23.2)$$

We will therefore develop a test to determine whether there is statistical evidence that (23.2) is false.

Estimating the Probabilities

In principle, given n total sample elements, the number of elements in each of the $r \cdot c$ cells follows a multinomial distribution with $r \cdot c - 1$ independent probabilities p_{ij} . (Recall that the sum over all probabilities must equal 1, so there are one fewer than $r \cdot c$ independently selectable parameters.)

However, if we assume $p_{ij} = p_{i\cdot}p_{\cdot j}$, then the multinomial distribution only depends on the $r - 1 + c - 1$ parameters $p_{i\cdot}$ and $p_{\cdot j}$. We will exploit this for our test.

Natural estimates for the row and column probabilities are

$$\widehat{p_{i\cdot}} = \frac{n_{i\cdot}}{n}, \quad \widehat{p_{\cdot j}} = \frac{n_{\cdot j}}{n}$$

so if (23.2) is assumed,

$$\widehat{p_{ij}} = \widehat{p_{i\cdot}}\widehat{p_{\cdot j}} = \frac{n_{i\cdot}n_{\cdot j}}{n^2}$$

Chi-Squared Test for Independence

Hence, if (23.2) is assumed, the expected number of elements in the (i, j) th cell is

$$E_{ij} = n \cdot \widehat{p}_{ij} = \frac{n_i \cdot n_j}{n}$$

We can now compare the observed frequencies O_{ij} in the (i, j) th cell to the expected frequencies E_{ij} . We will again use the Pearson statistic

$$\chi^2_{(r-1)(c-1)} = \sum_{i=1}^r \sum_{j=1}^c \frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

which follows a chi-squared distribution with

$$k - 1 - m = rc - 1 - (r + c - 2) = rc - r - c + 1 = (r - 1)(c - 1)$$

degrees of freedom. We reject H_0 if the value of $\chi^2_{(r-1)(c-1)}$ exceeds the critical value of the corresponding chi-squared distribution.

Testing for Independence

23.14. Example. A company has to choose among three pension plans. Management wishes to know whether the preference for plans is independent of job classification and wants to use $\alpha = 0.05$. The opinions of a random sample of 500 employees are shown below.

	Plan 1	Plan 2	Plan 3	Totals
Salaried Workers	160	140	40	$n_{1\cdot} = 340$
Hourly Workers	40	60	60	$n_{2\cdot} = 160$
Totals	$n_{\cdot 1} = 200$	$n_{\cdot 2} = 200$	$n_{\cdot 3} = 100$	$n = 500$

We want to test

H_0 : there is no dependence between job classification and plan preference

Testing for Independence

We calculate the expected frequencies assuming that H_0 is true:

$$\begin{aligned}E_{11} &= \frac{200 \cdot 340}{500}, & E_{12} &= \frac{200 \cdot 340}{500}, & E_{13} &= \frac{100 \cdot 340}{500} \\E_{21} &= \frac{200 \cdot 160}{500}, & E_{22} &= \frac{200 \cdot 160}{500}, & E_{23} &= \frac{100 \cdot 160}{500}\end{aligned}$$

It is now a simple matter to calculate the value of the statistic

$$\chi^2_{(2-1)(3-1)} = \chi^2_2 = \sum_{i=1}^2 \sum_{j=1}^3 \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = 49.63.$$

Since the statistic follows a chi-squared distribution with 2 degrees of freedom and we want $\alpha = 0.05$, we compare this value with $\chi^2_{0.05, 2} = 5.99$. As $49.63 > 5.99$, we may reject H_0 . There is evidence that the pension plan preference is not independent of job classification.

Comparing Proportions

Finally, we note that a very similar, though subtly different approach can be taken when comparing multiple proportions. Suppose that we would like to compare the proportions students with little, average or much nightly sleep among the JI ECE, JI ME, SJTU EE and SJTU ME majors. We choose to randomly select (based on student IDs) $n_{1\cdot}$, $n_{2\cdot}$, $n_{3\cdot}$ and $n_{4\cdot}$ students, respectively, from each of these majors.

We obtain the following contingency table:

	< 6h sleep	6–9h sleep	> 9h sleep	
JI ECE	n_{11}	n_{12}	n_{13}	$n_{1\cdot}$ (fixed)
JI ME	n_{21}	n_{22}	n_{23}	$n_{2\cdot}$ (fixed)
SJTU EE	n_{31}	n_{32}	n_{33}	$n_{3\cdot}$ (fixed)
SJTU ME	n_{41}	n_{42}	n_{43}	$n_{4\cdot}$ (fixed)
	$n_{\cdot 1}$	$n_{\cdot 2}$	$n_{\cdot 3}$	n (fixed)

Comparing Proportions

We again suppose that the number of objects in each cell is governed by the multinomial distribution. However, since the row totals are now fixed, only the number of objects in the first $c - 1$ columns can be independently chosen, so we have a total of $r \cdot (c - 1)$ independent cells.

It is useful to rewrite the above table in terms of proportions:

	< 6h sleep	6–9h sleep	> 9h sleep	
JI ECE	p_{11}	p_{12}	p_{13}	$p_{1\cdot} = 1$ (fixed)
JI ME	p_{21}	p_{22}	p_{23}	$p_{2\cdot} = 1$ (fixed)
SJTU EE	p_{31}	p_{32}	p_{33}	$p_{3\cdot} = 1$ (fixed)
SJTU ME	p_{41}	p_{42}	p_{43}	$p_{4\cdot} = 1$ (fixed)

We test

$$H_0 : \begin{cases} p_{11} = p_{21} = p_{31} = p_{41}, \\ p_{12} = p_{22} = p_{32} = p_{42}, \\ p_{13} = p_{23} = p_{33} = p_{43}. \end{cases}$$

Comparing Proportions

Supposing that H_0 is true, we have the common proportions

$$p_j := p_{1j} = p_{2j} = p_{3j} = p_{4j}, \quad j = 1, 2, 3,$$

where p_j is also equal to the proportion of all objects following into the j th column. An estimate for p_j is

$$\hat{p}_j = \frac{n_{\cdot j}}{n}, \quad j = 1, 2, 3. \quad (23.3)$$

and \hat{p}_j also serves as an estimator for all of the p_{ij} , $i = 1, \dots, 4$. If H_0 is true, the expected frequency in each cell is given by

$$E_{ij} = n_{i\cdot} \hat{p}_{ij} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

and we can again apply the Pearson chi-squared test.

Comparing Proportions

For the general case of r rows and c columns, the test

$$H_0: p_{1j} = p_{2j} = \cdots = p_{rj}, \quad j = 1, \dots, c.$$

is called a **test for homogeneity**. When using the Pearson statistic, note that the degrees of freedom are

$$r(c - 1) - (c - 1) = (r - 1)(c - 1)$$

where $r(c - 1)$ is the number of independent cells and $c - 1$ is the number of independent parameters \hat{p}_j that are estimated in (23.3).

We note that the tests for independence and for homogeneity appear absolutely the same in practice. That is not very surprising, since the null hypotheses are logically equivalent.

Simple Linear Regression I: Basic Model and Inferences

Linear Regression

Linear regression: modeling the dependency of two variables using a linear approach.

The term was originally used in the sense of **regression to the mean** in biology. It was observed that certain extreme values of biological features in members of a population are not necessarily passed on to descendants, but that the descendants' values of these features return to being closer to the mean.

However, after continued analysis of the concrete biological problem, the term “regression” came to be used much more generally.

Setting and Assumptions

We have:

- ▶ a **dependent variable Y** , which we will assume to be a random variable following a normal distribution. Y is often called the **response variable**.
- ▶ an **independent variable X** , which we can assume to either be a non-random parameter or a random variable measured precisely, without any error or uncertainty. X is often called the **predictor variable** or **regressor**.

We want to describe the behavior of Y as a function of the values of X , i.e., of $Y | x$. We will therefore assume that there exists a certain **model**.

For most of this discussion, we will take the point of view that x is not random while Y is a random variable following a normal distribution.

Simple Linear Regression Model

In this section, we assume that the mean $\mu_{Y|x}$ of $Y | x$ is given by

$$\mu_{Y|x} = \beta_0 + \beta_1 x \quad \text{for some } \beta_0, \beta_1 \in \mathbb{R}. \quad (24.1)$$

This is called a **simple linear regression** model with **model parameters** β_0 and β_1 .

Another way of writing this model is

$$Y | x = \beta_0 + \beta_1 x + E$$

where $E[E] = 0$.

Our goal is to find estimators

$$B_0 := \hat{\beta}_0 = \text{estimator for } \beta_0, \quad b_0 = \text{estimate for } \beta_0,$$

$$B_1 := \hat{\beta}_1 = \text{estimator for } \beta_1, \quad b_1 = \text{estimate for } \beta_1,$$

Residuals

We assume that we have a random sample of size n of pairs (X, Y) or (if we consider X to be a parameter and not a random variable) $(x, Y | x)$.

For short, we write

$$Y_i := Y | x_i, \quad i = 1, \dots, n,$$

so that we have a random sample $(x_1, Y_1), \dots, (x_n, Y_n)$.

For each measurement y_i there exists a number e_i , called the **residual**, such that

$$Y_i = b_0 + b_1 x_i + e_i.$$

Our goal is to determine b_0 and b_1 based on minimizing the residuals in a certain way.

Least-Squares Estimation

In 1805, Adrien Legendre published an approach for minimizing the residuals by letting

$$e_1^2 + e_2^2 + \cdots + e_n^2 \longrightarrow \text{minimum.}$$

Gauß published the same method (with a deeper analysis) in 1809 but claimed he had been using it since 1795. This set off a bitter priority dispute between Legendre and Gauß.

In a letter, Gauß notes that previously Laplace had been using the approach

$$|e_1| + |e_2| + \cdots + |e_n| \longrightarrow \text{minimum}$$

under the condition that $e_1 + e_2 + \cdots + e_n = 0$.

Gauß then extensively analyzed the **least-squares method**.



Adrien-Marie Legendre (1752-1833)
Bouilly, Julien-Leopold. (1820). Album de 73 Portraits-charge Aquarelles des Membres de l'Institut (watercolor portrait # 29). Bibliothèque de l'Institut de France.

Least Squares Estimation

Given a sample of size n , we define the **error sum of squares**

$$SS_E := e_1^2 + e_2^2 + \cdots + e_n^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i))^2.$$

Since we determine the estimators for β_0 and β_1 by minimizing this sum of squares, b_0 and b_1 are called **least-squares estimates**.

Assuming that $Y | x$ follows a normal distribution with variance σ^2 (independent of x) and mean $b_0 + b_1 x$, Gauß proved that the least-squares estimators have the smallest possible variance among all unbiased estimators for b_0 and b_1 .

The Normal Equations

We consider SS_E as a function of b_0 and b_1 and find the minimum by calculating the partial derivatives:

$$\frac{\partial SS_E}{\partial b_0} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i),$$

$$\frac{\partial SS_E}{\partial b_1} = -2 \sum_{i=1}^n (y_i - b_0 - b_1 x_i) x_i$$

Setting the derivatives equal to zero, we obtain the so-called **normal equations**

$$nb_0 + b_1 \sum_{i=1}^n x_i = \sum_{i=1}^n y_i, \quad b_0 \sum_{i=1}^n x_i + b_1 \sum_{i=1}^n x_i^2 = \sum_{i=1}^n x_i y_i$$

The Least Squares Estimates

These are linear equations for b_0 and b_1 , which may be easily solved:

$$b_1 = \frac{n \sum_{i=1}^n x_i y_i - \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)}{n \sum_{i=1}^n x_i^2 - \left(\sum_{i=1}^n x_i \right)^2}, \quad (24.2a)$$

$$b_0 = \frac{1}{n} \sum_{i=1}^n y_i - b_1 \cdot \frac{1}{n} \sum_{i=1}^n x_i. \quad (24.2b)$$

Although these formulas are straightforward for explicit calculations, it is worth re-writing them a little.

We will use the usual notation

$$\bar{x} = \frac{1}{n} \sum_{i=1}^n x_i \quad \text{and} \quad \bar{y} = \frac{1}{n} \sum_{i=1}^n y_i.$$

The Least Squares Estimates

Then it is easy to see that

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

For short, we will write

$$\begin{aligned}S_{xx} &:= \sum_{i=1}^n (x_i - \bar{x})^2, & S_{yy} &:= \sum_{i=1}^n (y_i - \bar{y})^2, \\ S_{xy} &:= \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}).\end{aligned}$$

The Least Squares Estimates

Then we can write

$$b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{S_{xy}}{S_{xx}}.$$

24.1. Example. Since humidity influences evaporation, the solvent balance of water-reducible paints during spray-out is affected by humidity. A controlled study is conducted to examine the relationship between humidity (X) and the extent of solvent evaporation (Y).

Knowledge of this relationship will be useful in that it will allow the painter to adjust his or her spray gun setting to account for humidity.

Linear Regression

The following data are obtained:

Observation	x	y	Observation	x	y
1	35.3	11.0	14	39.1	9.6
2	27.7	11.1	15	46.8	10.9
3	30.8	12.5	16	48.5	9.6
4	58.8	8.4	17	59.3	10.1
5	61.4	9.3	18	70.0	8.1
6	71.3	8.7	19	70.0	6.8
7	74.4	6.4	20	74.4	8.9
8	76.7	8.5	21	72.1	7.7
9	70.7	7.8	22	58.1	8.5
10	57.5	9.1	23	44.6	8.9
11	46.4	8.2	24	33.4	10.4
12	28.9	12.2	25	28.6	11.1
13	28.1	11.9			

Here x is the observed relative humidity (in %), y is the observed solvent evaporation (in %).

Linear Regression

We obtain

$$n = 25,$$

$$\sum_{i=1}^n x_i = 1312.9,$$

$$\sum_{i=1}^n y_i = 235.70,$$

$$\sum_{i=1}^n x_i^2 = 76193.7,$$

$$\sum_{i=1}^n y_i^2 = 2286.07,$$

$$\sum_{i=1}^n x_i y_i = 11802.2$$

Then, using the formulas (24.2), we have

$$b_1 = \hat{\beta}_1 = -0.0795,$$

$$b_0 = \hat{\beta}_0 = 13.6.$$

Hence the estimated regression equation is

$$\hat{\mu}_{Y|x} = 13.6 - 0.0795x.$$

For example, the mean solvent evaporation at 50% relative humidity is estimated to be 9.63%.

Linear Regression with Mathematica

We use the data from Example 24.1 to illustrate how linear regression is implemented:

```
data := {{35.3, 11.0}, {27.7, 11.1}, {30.8, 12.5}, {58.8, 8.4},  
{61.4, 9.3}, {71.3, 8.7}, {74.4, 6.4}, {76.7, 8.5},  
{70.7, 7.8}, {57.5, 9.1}, {46.4, 8.2}, {28.9, 12.2},  
{28.1, 11.9}, {39.1, 9.6}, {46.8, 10.9}, {48.5, 9.6},  
{59.3, 10.1}, {70.0, 8.1}, {70.0, 6.8}, {74.4, 8.9},  
{72.1, 7.7}, {58.1, 8.5}, {44.6, 8.9}, {33.4, 10.4},  
{28.6, 11.1}};  
model = LinearModelFit[data, x, x]
```

FittedModel [13.6013 - 0.0794677 x]

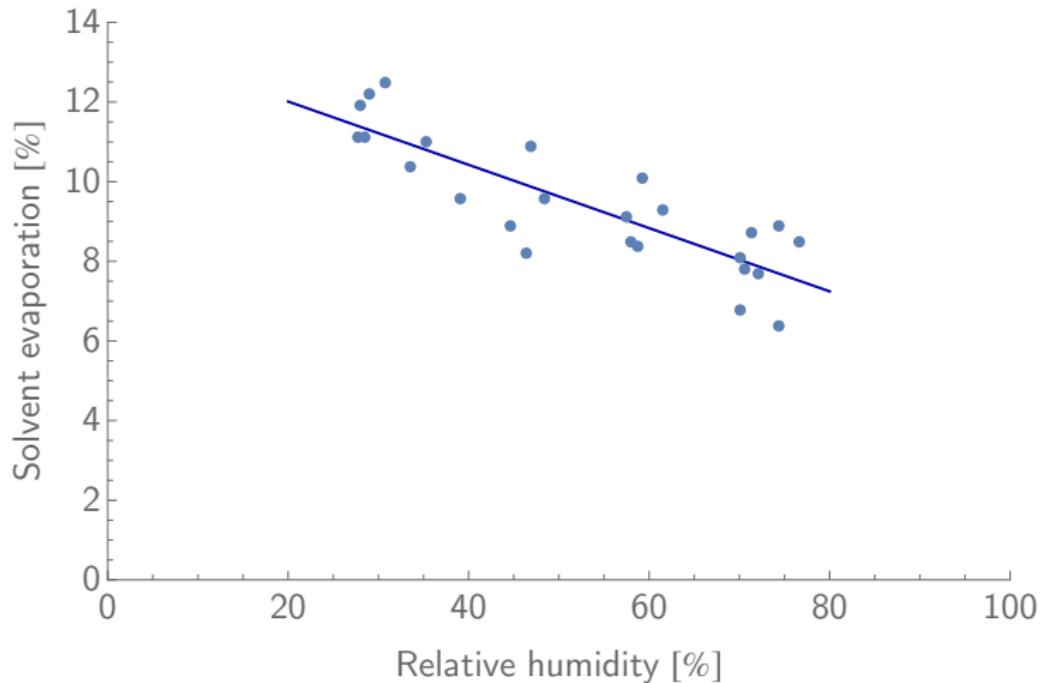
Data is entered as a list of pairs (x_i, y_i) and the **LinearModelFit** command takes as its arguments the data, the model (here: a linear model in x) and the name of the variable (x).



Linear Regression with Mathematica

```
model["BestFit"]
```

$$13.6013 - 0.0794677 x$$



Model Assumptions and Random Samples

24.2. Model Assumption.

- (i) For each value of x , the random variable $Y | x$ follows a normal distribution with variance σ^2 and mean $\mu_{Y|x} = \beta_0 + \beta_1 x$.
- (ii) The random variables $Y | x_1$ and $Y | x_2$ are independent if $x_1 \neq x_2$.

A random sample of size n consists of n pairs (x_i, Y_i) , $i = 1, \dots, n$, where the random variables $Y_i = Y | x_i$ are i.i.d. normal with variance σ^2 and mean $\mu_{Y|x_i} = \beta_0 + \beta_1 x_i$.

24.3. Remark. We do not require that $x_i \neq x_j$. The random sample may contain more than a single measurement of $Y | x_i$. All the x_i are treated in the same way, e.g., when calculating \bar{x} .

Distribution of the Least Squares Estimators

24.4. Theorem. Given a random sample of $Y | x$ of size n , the statistics

$$\frac{B_1 - \beta_1}{\sigma / \sqrt{\sum (x_i - \bar{x})^2}} \quad \text{and} \quad \frac{B_0 - \beta_0}{\sigma \sqrt{\frac{\sum x_i^2}{n \sum (x_i - \bar{x})^2}}}$$

follow a standard normal distribution.

In particular, B_0 and B_1 are unbiased estimators.

Distribution of the Least Squares Estimators

Proof.

We will prove the statement for the slope only. Since

$$\sum_{i=1}^n (x_i - \bar{x}) = 0$$

we may write

$$B_1 = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x})(Y_i - \bar{Y}) = \frac{1}{S_{xx}} \sum_{i=1}^n (x_i - \bar{x}) Y_i \quad (24.3)$$

Now B_1 is a linear combination of the i.i.d. normally distributed Y_i , so B_1 itself follows a normal distribution.

It remains to show that B_1 has mean β_1 and variance $\sigma^2 / \sum_{i=1}^n (x_i - \bar{x})^2$.

Distribution of the Least Squares Estimators

Proof (continued).

$$\begin{aligned} E[B_1] &= E\left[\sum_{i=1}^n \frac{x_i - \bar{x}}{S_{xx}} Y_i\right] = \sum_{i=1}^n \frac{x_i - \bar{x}}{S_{xx}} E[Y_i] \\ &= \sum_{i=1}^n \frac{(x_i - \bar{x})(\beta_0 + \beta_1 x_i)}{\sum_{i=1}^n (x_i - \bar{x})^2} \\ &= \frac{\beta_0}{\sum_{i=1}^n (x_i - \bar{x})^2} \underbrace{\sum_{i=1}^n (x_i - \bar{x})}_{=0} + \beta_1 \underbrace{\frac{\sum_{i=1}^n x_i(x_i - \bar{x})}{\sum_{i=1}^n (x_i - \bar{x})^2}}_{=1} \\ &= \beta_1. \end{aligned}$$

Distribution of Least Squares Estimators

Proof (continued).

Similarly,

$$\begin{aligned}\text{Var } B_1 &= \text{Var} \left(\sum_{i=1}^n \frac{x_i - \bar{x}}{S_{xx}} Y_i \right) = \sum_{i=1}^n \frac{(x_i - \bar{x})^2}{S_{xx}^2} \text{Var } Y_i \\ &= \frac{\sigma^2}{\left(\sum_{j=1}^n (x_j - \bar{x})^2 \right)^2} \sum_{i=1}^n (x_i - \bar{x})^2 \\ &= \frac{\sigma^2}{\sum_{i=1}^n (x_i - \bar{x})^2}.\end{aligned}$$

The proof of the corresponding statement for the estimator B_0 is completely analogous. □

Least Squares Estimator for the Variance

The variance σ^2 of $Y | x$ is assumed to be the same for all values of x . To estimate it, we use the error sum of squares,

$$S^2 := \frac{SS_E}{n-2} = \frac{1}{n-2} \sum_{i=1}^n (y_i - \hat{\mu}_{Y|x_i})^2 \quad (24.4)$$

It turns out that this estimator is unbiased for σ^2 and in fact

$$(n-2)S^2/\sigma^2 = \frac{SS_E}{\sigma^2}$$

follows a chi-squared distribution with $n - 2$ degrees of freedom.

Furthermore, it can be shown that S^2 is independent of B_0 and B_1 .

(Analogously to the statement that the sample mean is independent of the sample variance, which we proved using the Helmert transformation.)

Inferences on the Slope and the Intercept

Therefore,

$$\frac{(B_1 - \beta_1)/(\sigma/\sqrt{S_{xx}})}{\sqrt{(n-2)S^2/[\sigma^2(n-2)]}} = \frac{B_1 - \beta_1}{S/\sqrt{S_{xx}}}$$

follows a T -distribution with $n - 2$ degrees of freedom.

The same is true for

$$\frac{(B_0 - \beta_0)/(\sigma\sqrt{\sum x_k^2}/\sqrt{nS_{xx}})}{\sqrt{(n-2)S^2/[\sigma^2(n-2)]}} = \frac{B_0 - \beta_0}{S\sqrt{\sum x_k^2}/\sqrt{nS_{xx}}}.$$

It follows immediately that we have $100(1 - \alpha)\%$ confidence intervals

$$B_1 \pm t_{\alpha/2, n-2} \frac{S}{\sqrt{S_{xx}}},$$

$$B_0 \pm t_{\alpha/2, n-2} \frac{S\sqrt{\sum x_i^2}}{\sqrt{nS_{xx}}}$$

for β_1 and β_0 , respectively.

Practical Calculations

In practice, to simplify calculations using a calculator, we may use the following relations:

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

$$S_{yy} = \sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2,$$

$$S_{xy} = \sum_{i=1}^n (x_i - \bar{x})(y_i - \bar{y}) = \sum_{i=1}^n x_i y_i - \frac{1}{n} \left(\sum_{i=1}^n x_i \right) \left(\sum_{i=1}^n y_i \right)$$

and

$$b_0 = \bar{y} - b_1 \bar{x}, \quad b_1 = \frac{S_{xy}}{S_{xx}}, \quad SS_E = S_{yy} - b_1 S_{xy}. \quad (24.5)$$

Confidence Intervals for Slope and Intercept

24.5. Example. We return to Example 24.1 of solvent evaporation in spray painting. Recall that we obtained the point estimate for the regression line

$$\hat{\mu}_{Y|x} = 13.6 - 0.0795x.$$

Based on the previously calculated

$$\begin{aligned} n &= 25, & \sum x_i &= 1312.9, & \sum y_i &= 235.70, \\ \sum x_i^2 &= 76193.7, & \sum y_i^2 &= 2286.07, & \sum x_i y_i &= 11802.2 \end{aligned}$$

we obtain

$$S_{xx} = 7245.47, \quad S_{yy} = 63.89, \quad S_{xy} = -575.781$$

and

$$SS_E = S_{yy} - b_1 S_{xy} = 18.13, \quad s^2 = SS_E / (n - 2) = 0.79$$

Confidence Intervals for Slope and Intercept

A 95% confidence interval for the slope of the regression line is given by

$$\begin{aligned} b_1 \pm t_{0.025, 23}s/\sqrt{S_{xx}} &= -0.0795 \pm \frac{2.0687 \cdot 0.888}{85.12} \\ &= -0.0795 \pm 0.0215 \end{aligned}$$

and a 95% confidence interval for the intercept is given by

$$b_0 \pm t_{0.025, 23}s\sqrt{\sum x_i^2}/\sqrt{nS_{xx}} = 13.6 \pm 1.19$$

These confidence intervals can also be obtained with Mathematica:

```
model["ParameterConfidenceIntervals", ConfidenceLevel → 0.95]
{{12.41, 14.7927}, {-0.101047, -0.0578881}}
```

Tests for Regression Parameters

Of course, we may also perform hypothesis tests (Fisher or Neyman-Pearson) on the model parameters. For example, we may test

$$H_0: \beta_0 = \beta_0^0$$

and

$$H_0: \beta_1 = \beta_1^0$$

for null values β_0^0 and β_1^0 of the intercept and slope, respectively.

An important special case is following:

We say that a regression is **significant** if there is statistical evidence that the slope $\beta_1 \neq 0$.

Test for Significance of Regression

24.6. Test for Significance of Regression. Let $(x_i, Y | x_i), i = 1, \dots, n$ be a random sample from $Y | x$. We reject

$$H_0: \beta_1 = 0$$

at significance level α if the statistic

$$T_{n-2} = \frac{B_1}{S/\sqrt{S_{xx}}}.$$

satisfies $|T_{n-2}| > t_{\alpha/2, n-2}$.

Significance of Regression

24.7. Example. We return to Example 24.1 of solvent evaporation in spray painting. Recall that we obtained the point estimate for the regression line

$$\hat{\mu}_{Y|x} = 13.64 - 0.08x.$$

We now test whether the regression is significant. from the previously calculated data, we have

$$t_{23} = \frac{b_1}{s/\sqrt{S_{xx}}} = -7.62.$$

We find that $P[T_{23} \leq -7.62] < 0.0005$.

Since this is a two-tailed test, $P < 2 \cdot 0.0005 = 0.001$.

Hence, we are able to reject H_0 . There is no evidence that the regression is not significant.

Properties of the Estimator for the Mean

We now turn to the actual estimated mean, $\mu_{Y|x}$. The least-squares estimators give

$$\hat{\mu}_{Y|x} = B_0 + B_1 x.$$

Since B_0 and B_1 are unbiased estimators for β_0 and β_1 it follows immediately that $\hat{\mu}_{Y|x}$ is unbiased for $\mu_{Y|x}$.

We may write

$$\hat{\mu}_{Y|x} = B_0 + B_1 x = \bar{Y} - B_1 \bar{x} + B_1 x = \bar{Y} + B_1(x - \bar{x}).$$

Since B_1 is a linear combination of the Y_i (see (24.3)), this implies that $\hat{\mu}_{Y|x}$ is also a linear combination of the Y_i . The Y_i are assumed independent and normally distributed, so we see that $\hat{\mu}_{Y|x}$ follows a normal distribution.

Distribution of the Estimated Mean

Since $\text{Cov}(\bar{Y}, B_1) = 0$ (see assignments),

$$\begin{aligned}\text{Var}[\hat{\mu}_{Y|x}] &= \text{Var}[\bar{Y} + (x - \bar{x})B_1] \\ &= \text{Var}[\bar{Y}] + (x - \bar{x})^2 \text{Var } B_1 \\ &= \frac{\sigma^2}{n} + \frac{(x - \bar{x})^2 \sigma^2}{S_{xx}}.\end{aligned}$$

In conclusion,

$$\frac{\hat{\mu}_{Y|x} - \mu_{Y|x}}{\sigma \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}}$$

follows a standard-normal distribution.

Confidence Interval the Estimated Mean

Using our estimator for the variance as before, we see that

$$\frac{\hat{\mu}_{Y|x} - \mu_{Y|x}}{S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}}$$

follows a T distribution with $n - 2$ degrees of freedom.

Based on this, we may make inferences on the value of the mean of $Y | x$.

For example, we obtain the following $100(1 - \alpha)\%$ confidence interval for $\mu_{Y|x}$:

$$\hat{\mu}_{Y|x} \pm t_{\alpha/2, n-2} S \sqrt{\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \quad (24.6)$$

Simple Linear Regression II: Predictions and Model Analysis

Inferences about a Single Predicted Value

We are interested in finding an “estimate” (guess) or a ***prediction*** for the value of the random variable $Y | x$. Note the essential difference:

- ▶ An ***estimate*** is a statistical statement on the value of an unknown, but fixed, population parameter.
- ▶ A ***prediction*** is a statistical statement on the value of an essentially random quantity.

We define a $100(1 - \alpha)\%$ prediction interval $[L_1, L_2]$ for a random variable X by

$$P[L_1 \leq X \leq L_2] = 1 - \alpha.$$

As a ***predictor*** $\widehat{Y | x}$ for the value of $Y | x$ we use the estimator for the mean, i.e., we set

$$\widehat{Y | x} = \hat{\mu}_{Y|x} = B_0 + B_1 x.$$

In order to find a prediction interval, we need to analyze the distribution of $\widehat{Y | x}$.

Inferences about a Single Predicted Value

Recall that $\hat{\mu}_{Y|x}$ follows a normal distribution with mean $\mu_{Y|x}$ and variance $\left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)\sigma^2$. Furthermore, $Y | x$ is normally distributed with mean $\mu_{Y|x}$ and variance σ^2 .

Hence $\widehat{Y|x} - Y|x$ is normally distributed and, furthermore,

$$E[\widehat{Y|x} - Y|x] = \mu_{Y|x} - \mu_{Y|x} = 0,$$

$$\text{Var}[\widehat{Y|x} - Y|x] = \left(\frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)\sigma^2 + \sigma^2 = \left(1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}\right)\sigma^2.$$

Thus, after standardizing and dividing by S/σ we obtain the T_{n-2} random variable

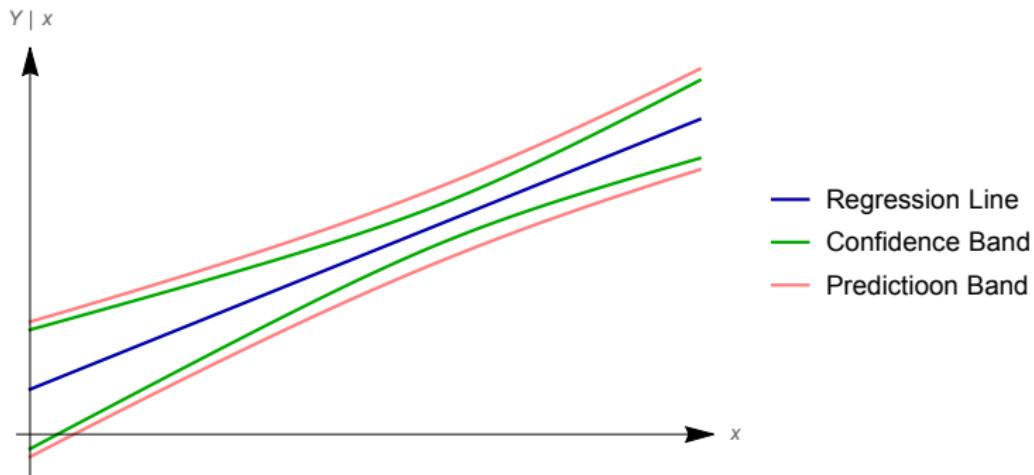
$$T_{n-2} = \frac{\widehat{Y|x} - Y|x}{S\sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}}}$$

Inferences about a Single Predicted Value

We thus obtain the following $100(1 - \alpha)\%$ prediction interval for $Y | x$:

$$\widehat{Y} | x \pm t_{\alpha/2, n-2} S \sqrt{1 + \frac{1}{n} + \frac{(x - \bar{x})^2}{S_{xx}}} \quad (25.1)$$

The limits of the confidence interval (24.6) and the prediction interval (25.1), plotted as functions of x , are commonly called **confidence bands** and **prediction bands** for the regression.



Confidence and Prediction Intervals

25.1. Example. Continuing with the data from Example 24.1, Mathematica gives confidence bands (24.6) for the estimated mean as

```
conf = model["MeanPredictionBands", ConfidenceLevel → 0.95]
```

$$\left\{ 13.6013 - 0.0794677 x - 2.06866 \sqrt{0.331656 - 0.0114296 x + 0.00010882 x^2}, \right. \\ \left. 13.6013 - 0.0794677 x + 2.06866 \sqrt{0.331656 - 0.0114296 x + 0.00010882 x^2} \right\}$$

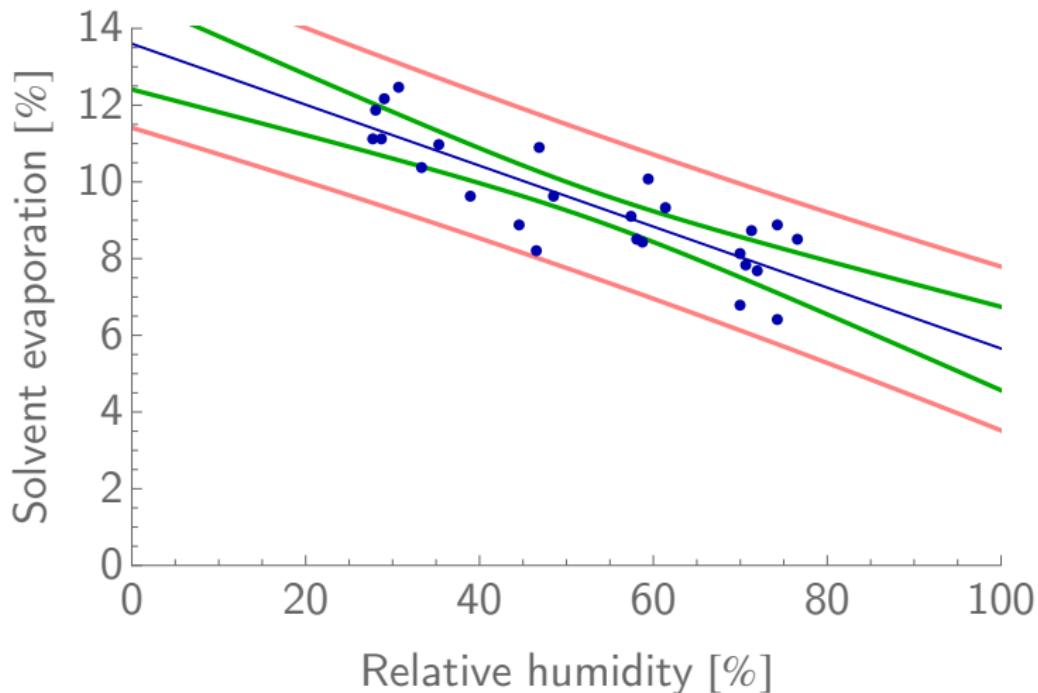
Prediction bands (25.1) are given by

```
pred = model["SinglePredictionBands", ConfidenceLevel → 0.95]
```

$$\left\{ 13.6013 - 0.0794677 x - 2.06866 \sqrt{1.12011 - 0.0114296 x + 0.00010882 x^2}, \right. \\ \left. 13.6013 - 0.0794677 x + 2.06866 \sqrt{1.12011 - 0.0114296 x + 0.00010882 x^2} \right\}$$

★ Confidence and Prediction Intervals

Below, the prediction bands are shown in red, while the confidence bands for the estimated mean are green:



Analysis of the Model

Achievements so far:

- ▶ Inferences on model parameters β_0, β_1 .
- ▶ Inferences on estimated mean $\hat{\mu}_{Y|x}$.
- ▶ Prediction for $Y | x$.

But is our linear model actually appropriate?

Crucial Quantities:

- ▶ The total variation of the response variable,

$$SS_T = S_{yy} = \sum_{i=1}^n (Y_i - \bar{Y})^2.$$

We will also call this the ***Total Sum of Squares***. It represents the variation of Y regardless of any model.

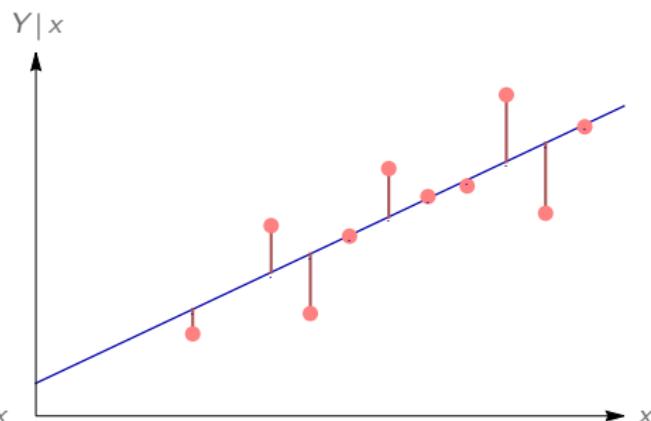
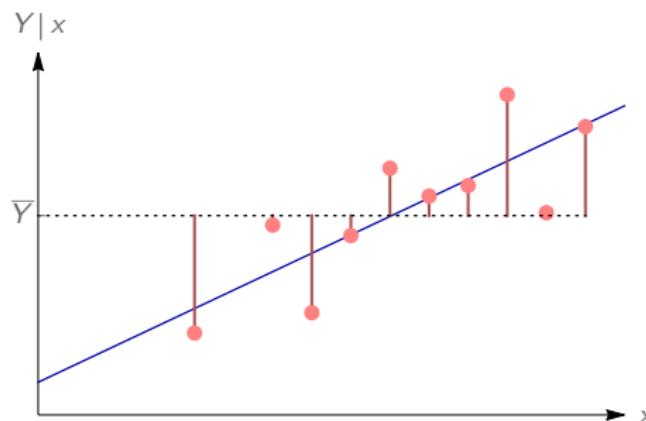
Analysis of the Model

Crucial Quantities:

- The **Error Sum of Squares**

$$SS_E = \sum_{i=1}^n (Y_i - (b_0 + b_1 x))^2.$$

It represents the variation of Y that remains after we have applied the model.



Coefficient of Determination

Of course,

$$SS_E \leq SS_T$$

and we define the the ***coefficient of determination***

$$R^2 := \frac{SS_T - SS_E}{SS_T}. \quad (25.2)$$

Sometimes, one reports $R^2 \cdot 100\%$.

The coefficient R^2 expresses the ***proportion of the total variation in Y that is explained by the linear model.***

Connection to Correlation

Recall from (24.5) that

$$SS_E = S_{yy} - B_1 S_{xy} = S_{yy} - \frac{S_{xy}^2}{S_{xx}},$$

so that

$$R^2 = \frac{SS_T - SS_E}{SS_T} = \frac{S_{xy}^2}{S_{xx} S_{yy}}.$$

The right-hand side is exactly the square of the estimator (22.1) for the correlation coefficient ρ_{XY} .

Since the correlation ρ_{XY} measures the linearity of the relationship between X and Y , this is not surprising.

Connection to Significance of Regression

The statistic that we have used in the Test for Significance of regression 24.6 is

$$\frac{B_1}{\sqrt{S^2/S_{xx}}} = \frac{S_{xy}/S_{xx}}{\sqrt{SS_E / [(n - 2)S_{xx}]}} \quad (25.3)$$

$$= \frac{R}{\sqrt{1 - R^2}} \sqrt{n - 2}, \quad (25.4)$$

and we can see that is expressible entirely using the coefficient R^2 .

Hence, R^2 alone includes enough information to conduct the test for significance of regression.

Test for Correlation

Conversely, we can adapt the above discussion to perform a two-sided Fisher test for a vanishing correlation in a bivariate normal distribution:

25.2. Test for Correlation. Let (X, Y) follow a bivariate normal distribution with correlation coefficient $\rho \in (-1, 1)$. Let R be the estimator (22.1) for ρ . Then

$$H_0: \rho = 0$$

is rejected at significance level α if

$$\left| \frac{R\sqrt{n-2}}{\sqrt{1-R^2}} \right| > t_{\alpha/2, n-2}.$$

Lack-of-Fit and Pure Error

Problem:

- ▶ R^2 measures how much of the total variation is explained by the linear model.
- ▶ If R^2 is not large, then the model does not explain a significant amount of the fluctuation of the measured values y_i .
- ▶ In short, SS_E is large.

Why could SS_E be large?

- ▶ Either σ^2 is very large (**pure error**)
- ▶ or the model is wrong. (**lack-of-fit error**)

To tell which of the two predominates, we need to be able to take **repeated measurements** of $Y | x_i$ for the same value of x_i .

Repeated Measurements

We can directly measure pure error (due to σ^2) if we have ***repeated measurements*** available. That is, at one or more points x_i , $i = 1, \dots, k$, we have at least two observations on Y .

Let Y_{ij} denote the j th observation of $Y | x_i$, where $j = 1, \dots, n_i$.

The total number of observations is

$$n = n_1 + n_2 + \cdots + n_k = \sum_{i=1}^k n_i.$$

Recall: Repeated measurements are treated just like any other measurements in regression analysis.

Internal Sum of Squares

For each $i = 1, \dots, k$ we can view $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$ as a random sample of size n_i of the random variable $Y_i = Y | x_i$.

An unbiased estimator for $\mu_{Y|x_i}$ is the sample mean

$$\bar{Y}_i = \frac{1}{n_i} \sum_{j=1}^{n_i} Y_{ij}.$$

The statistic

$$\sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

measures the natural variability of $Y | x_i$ and is called an **internal sum of squares**.

Error Sum of Squares (Pure Error)

By summing over all internal sums of squares we obtain the **error sum of squares due to pure error**,

$$\begin{aligned} \text{SS}_{E;\text{pe}} &:= \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2 \quad (25.5) \\ &= \sum_{i=1}^k \sum_{j=1}^{n_i} Y_{ij}^2 - \sum_{i=1}^k \frac{1}{n_i} \left(\sum_{j=1}^{n_i} Y_{ij} \right)^2 \end{aligned}$$

It is not difficult to see that

$$\frac{1}{\sigma^2} \text{SS}_{E;\text{pe}} = \sum_{i=1}^k \frac{1}{\sigma^2} \sum_{j=1}^{n_i} (Y_{ij} - \bar{Y}_i)^2$$

follows a chi-squared distribution with $n - k$ degrees of freedom.

Error Sum of Squares

Note that $SS_{E;f} \leq SS_E$ since

$$SS_E = \sum_{i=1}^k \sum_{j=1}^{n_i} (Y_{ij} - (b_0 + b_1 x_i))^2$$

and in general

$$\sum_{j=1}^{n_i} (Y_{ij} - z)^2$$

is minimized if $z = \bar{Y}_i$.

Error Sum of Squares (Lack of Fit)

We therefore define the ***error sum of squares due to lack of fit*** by

$$SS_{E;lf} := SS_E - SS_{E;pe}.$$

Since $SS_E = SS_{E;pe} + SS_{E;lf}$, it seems reasonable that

$$\frac{1}{\sigma^2} SS_{E;lf}$$

might follow a chi-squared distribution with

$$(n - 2) - (n - k) = k - 2$$

degrees of freedom.

In fact, it can be shown that this is true and that $SS_{E;lf}$ is actually a sum of squares.

Testing for Lack of Fit

25.3. Test for Lack of Fit. Let x_1, \dots, x_k be regressors and $Y_{i1}, Y_{i2}, \dots, Y_{in_i}$, $i = 1, \dots, k$, the measured responses at each of the regressors. Let $SS_{E;pe}$ and $SS_{E;lf}$ be the pure error and lack-of-fit sums of squares for a linear regression model. Then

H_0 : the linear regression model is appropriate

is rejected at significance level α if the test statistic

$$F_{k-2,n-k} = \frac{SS_{E;lf}/(k-2)}{SS_{E;pe}/(n-k)}$$

satisfies $F_{k-2,n-k} > f_{\alpha,k-2,n-k}$.

Testing for Lack of Fit

25.4. Example. Consider these data on X , the temperature, in degrees centigrade, at which a chemical reaction is conducted, and Y , the percentage yield obtained:

x_i	30	40	50	60	70
Y_{i1}	13.7	15.5	18.5	17.7	15.0
Y_{i2}	14.0	16.0	20.0	18.1	15.6
Y_{i3}	14.6	17.0	21.1	18.5	16.5

Here $k = 5$, $n_1, \dots, n_5 = 3$ and $n = 15$. For $x_1 = 30$ we have $\bar{y}_1 = 14.1$ and the internal sum of squares

$$(13.7 - 14.1)^2 + (14.0 - 14.1)^2 + (14.6 - 14.1)^2 = 0.42$$

In the same way we calculate the other two internal sums of squares and obtain the pure error sum of squares

$$SS_{E;pe} = 6.453.$$

Testing for Lack of Fit

For our data we can calculate $S_{yy} = 66.6437$, $S_{xy} = 154$ and $b_1 = 0.051$.
The total error sum of squares is given by

$$SS_E = S_{yy} - b_1 S_{xy} = 58.583.$$

The lack-of-fit sum of squares is

$$SS_{E;lf} = SS_E - SS_{E;pe} = 52.13.$$

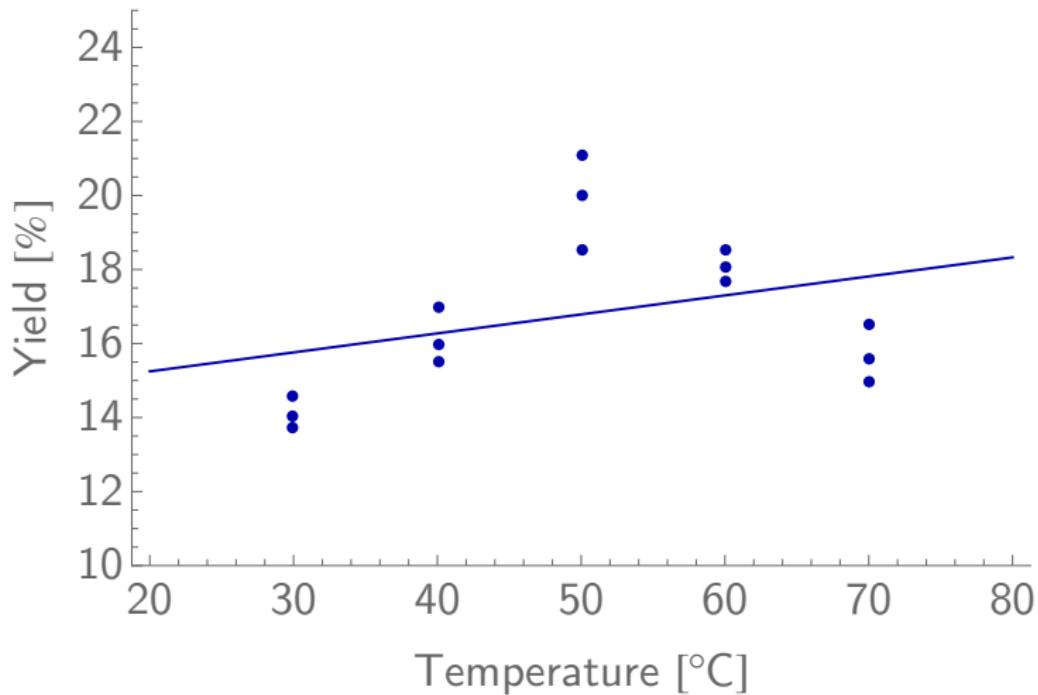
The observed value of the statistic

$$F_{k-2, n-k} = F_{3,10} = \frac{SS_{E;lf} / (k-2)}{SS_{E;pe} / (n-k)} = \frac{52.13/3}{6.453/10} = 26.928.$$

Based on the $F_{3,10}$ distribution, we can reject H_0 with $P < 0.05$
($f_{0.05,3,10} = 3.708$). There is evidence that a linear regression model is not appropriate.

Testing for Lack of Fit

It is clear from the graph below that the linear model is indeed not suitable:



Residual Analysis

The residuals e_i , $i = 1, \dots, n$ give important information on the model:

- ▶ Are they consistent with the assumption of equal variance σ^2 ?
- ▶ Are they consistent with the assumption of a normal distribution?
- ▶ Does the linear model seem appropriate?

Plotting the residuals vs. the values of x_i also shows potential gaps in the data.

Never extrapolate the regression model beyond the range of the regressors. Avoid leaving wide gaps in the range of the x_i .

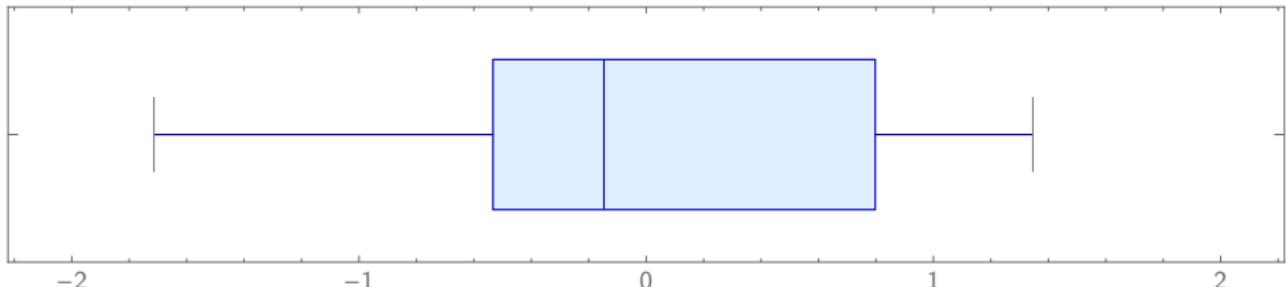
 Residual Analysis

25.5. Example. Continuing with the data from Example 24.1, Mathematica gives the residuals as follows

```
residuals = model["FitResiduals"]
```

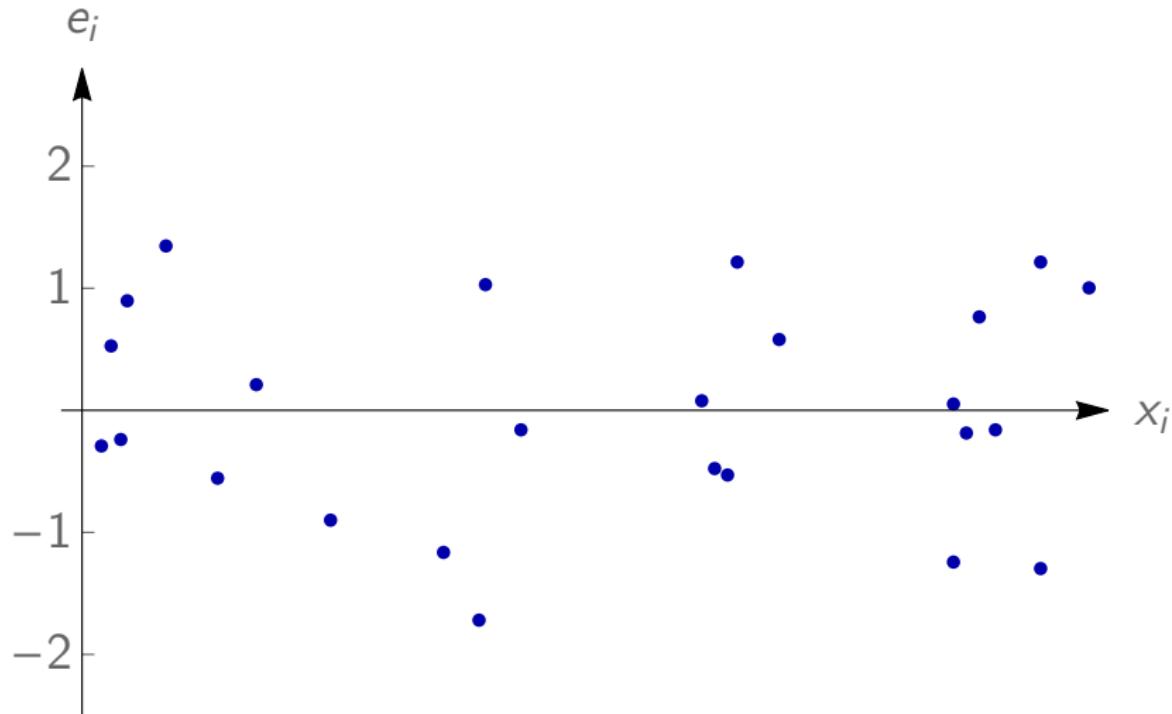
```
{0.203884, -0.300071, 1.34628, -0.528625, 0.577991, 0.764721,  
-1.28893, 0.993847, -0.182959, 0.0680671, -1.71402, 0.895291,  
0.531716, -0.894139, 1.01776, -0.147142, 1.21111, 0.0614135, -1.23859,  
1.21107, -0.171704, -0.484252, -1.15707, -0.547105, -0.22855}
```

A boxplot does not yield strong evidence against the normality assumption:



Residual Analysis

The residual plot does not show any obvious issues:



The Anscombe Quartet



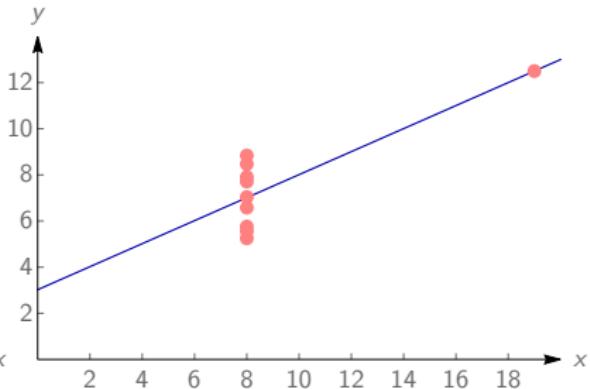
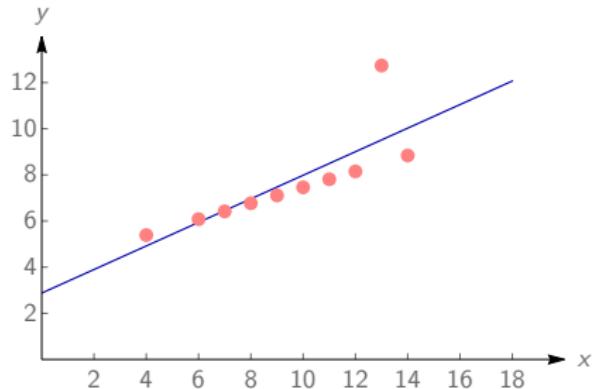
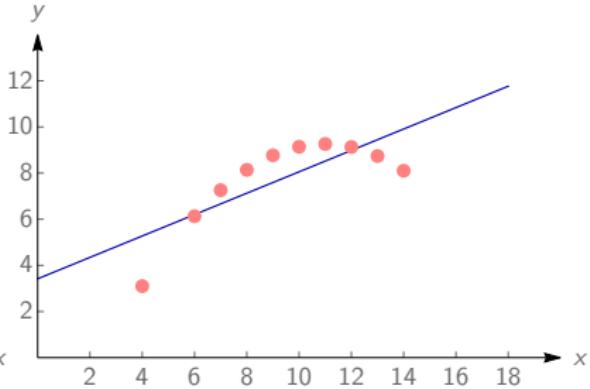
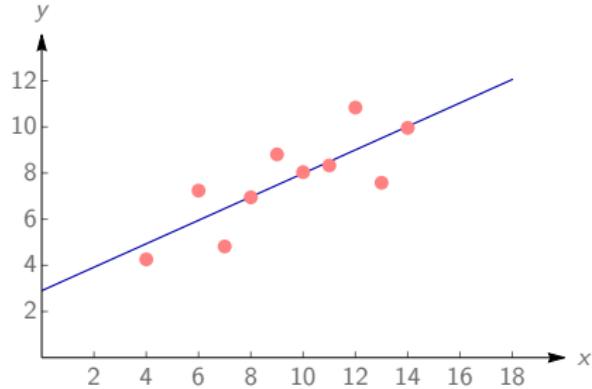
Francis J. Anscombe (1918-2001)
Boilly, Julien-Leopold. (1820). Yale
Bulletin and Calendar. November 2,
2001. Vol. 30, No. 9

- ▶ $\bar{x} = 9$,
- ▶ $\frac{1}{n-1} S_{xx} = 11$,
- ▶ $\bar{y} = 7.50$,
- ▶ $\frac{1}{n-1} S_{yy} = 4.122 \text{ or } 4.127$,
- ▶ $R^2 = 0.816$,
- ▶ $\hat{\mu}_{Y|x} = 3.00 + 0.500x$

Finally, this example by Francis Anscombe illustrates why it is always important to actually look at the data instead of relying on numerical quantities. In the following four graphs, the data all have these same statistics up to the precision given.

Literature: Anscombe, F. J. *Graphs in Statistical Analysis*. American Statistician. 27 (1): 17–21. (1973).

Plot the data!



Multiple Linear Regression I: Basic Model

More General Regression Models

Two Main Generalizations:

- The ***multilinear model*** with linear dependence on $p \in \mathbb{N}$ parameters X_1, \dots, X_p ,

$$\mu_{Y|x_1, \dots, x_p} = \beta_0 + \beta_1 x_1 + \dots + \beta_p x_p \quad (26.1)$$

and

- The ***polynomial model*** with dependence on a polynomial of degree p of a single parameter X ,

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2 + \dots + \beta_p x^p.$$

The Polynomial Model

A random sample of size n , $(x_i, Y | x_i)$, $i = 1, \dots, n$ is given. We write $Y_i := Y | x_i$ as usual.

Goal: Find b_0, \dots, b_p such that for

$$y_i = b_0 + b_1 x_i + \cdots + b_p x_i^p + e_i, \quad i = 1, \dots, n, \quad (26.2)$$

the sum of squares error

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_i + \cdots + b_p x_i^p))^2 \quad (26.3)$$

is minimized.

The Model Specification Matrix

To discuss the model

$$Y_i = Y \mid x_i = \beta_0 + \beta_1 x_i + \beta_2 x_i^2 + \cdots + \beta_p x_i^p + E_i, \quad (26.4)$$

it is convenient to adopt a matrix formalism. We define

$$\mathbf{Y} = \begin{pmatrix} Y_1 \\ \vdots \\ Y_n \end{pmatrix}, \quad \mathbf{X} = \begin{pmatrix} 1 & x_1 & x_1^2 & \dots & x_1^p \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 1 & x_n & x_n^2 & \dots & x_n^p \end{pmatrix}, \quad \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \vdots \\ \beta_p \end{pmatrix}, \quad \mathbf{E} = \begin{pmatrix} E_1 \\ \vdots \\ E_n \end{pmatrix}.$$

Then (26.4) can be written as

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}. \quad (26.5)$$

The matrix \mathbf{X} is called the **model specification matrix**. We see from (26.5) that the polynomial model is a **linear regression model**.

Polynomial Regression

Defining

$$\hat{\beta} = \mathbf{b} := \begin{pmatrix} b_0 \\ \vdots \\ b_p \end{pmatrix} \quad \text{and} \quad \mathbf{e} = \begin{pmatrix} e_1 \\ \vdots \\ e_n \end{pmatrix}$$

(26.2) becomes

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e},$$

where \mathbf{b} is chosen to minimize the error sum of squares

$$SS_E = \langle \mathbf{Y} - \mathbf{X}\mathbf{b}, \mathbf{Y} - \mathbf{X}\mathbf{b} \rangle = (\mathbf{Y} - \mathbf{X}\mathbf{b})^T(\mathbf{Y} - \mathbf{X}\mathbf{b}). \quad (26.6)$$

Here A^T denotes the transpose of a matrix A and $\langle a, b \rangle = \sum_{i=1}^n a_i b_i$ is the usual scalar product of two vectors $a, b \in \mathbb{R}^n$.

Minimizing the SS_E

Using the norm $\|a\| = \sqrt{\langle a, a \rangle}$, we write

$$\begin{aligned} SS_E &= \langle \mathbf{Y} - X\mathbf{b}, \mathbf{Y} - X\mathbf{b} \rangle \\ &= \|\mathbf{Y}\|^2 - 2\langle X\mathbf{b}, \mathbf{Y} \rangle + \|X\mathbf{b}\|^2. \end{aligned}$$

The minimum of the sum-of-squares error is found from

$$\nabla_{\mathbf{b}} SS_E = \begin{pmatrix} \frac{\partial SS_E}{\partial b_0} \\ \vdots \\ \frac{\partial SS_E}{\partial b_p} \end{pmatrix} = 0.$$

Hence, we need to solve

$$-2\nabla_{\mathbf{b}} \langle X\mathbf{b}, \mathbf{Y} \rangle + \nabla_{\mathbf{b}} \langle X\mathbf{b}, X\mathbf{b} \rangle = 0.$$

Minimizing the SS_E

We use that

$$\langle X\mathbf{b}, \mathbf{Y} \rangle = \langle \mathbf{b}, X^T \mathbf{Y} \rangle = \sum_{i=0}^p b_i (X^T \mathbf{Y})_{i+1}$$

to see

$$\frac{\partial}{\partial b_k} \langle X\mathbf{b}, \mathbf{Y} \rangle = (X^T \mathbf{Y})_{k+1}.$$

and hence

$$\nabla_{\mathbf{b}} \langle X\mathbf{b}, \mathbf{Y} \rangle = \begin{pmatrix} \frac{\partial \langle X\mathbf{b}, \mathbf{Y} \rangle}{\partial b_0} \\ \vdots \\ \frac{\partial \langle X\mathbf{b}, \mathbf{Y} \rangle}{\partial b_p} \end{pmatrix} = \begin{pmatrix} (X^T \mathbf{Y})_1 \\ \vdots \\ (X^T \mathbf{Y})_{p+1} \end{pmatrix} = X^T \mathbf{Y}.$$

It is also not difficult to show that

$$\nabla_{\mathbf{b}} \langle X\mathbf{b}, X\mathbf{b} \rangle = 2X^T X\mathbf{b}.$$

The Regression Coefficients

It follows that the stationary point is given by

$$(X^T X) \boldsymbol{b} = X^T \mathbf{Y}.$$

Since the entries of X are numerical, $X^T X$ will almost surely be invertible.

Then the regression coefficients are given by

$$\boldsymbol{b} = (X^T X)^{-1} X^T \mathbf{Y}.$$

Of course, this formulation can also be used for simple linear regression; this is just the case $p = 1$.

Since the values in X are numerical, practical calculations are best done using a computer.

A Polynomial Model

26.1. Example. A study is conducted to develop an equation by which the unit cost of producing a new drug (Y) can be predicted based on the number of units produced (X). The proposed model is

$$\mu_{Y|x} = \beta_0 + \beta_1 x + \beta_2 x^2.$$

The following data are available:

x	5	5	10	10	15	15	20	20	25	25
y	14.0	12.5	7.0	5.0	2.1	1.8	6.2	4.9	13.2	14.6

We will use Mathematica for our calculations. We first enter the data as a list of pairs:

```
data = {{5, 14}, {5, 12.5}, {10, 7.}, {10, 5.}, {15, 2.1},  
{15, 1.8}, {20, 6.2}, {20, 4.9}, {25, 13.2}, {25, 14.6}};
```

A Polynomial Model

We construct the model specification matrix X and the response vector y :

```
y = Transpose[data][[2]];
X = Transpose[Table[Function[x, x^k] /@
Transpose[data][[1]], {k, 0, 2}]];
{MatrixForm[X], MatrixForm[y]}
```

$$\left\{ \begin{array}{ccc|c} 1 & 5 & 25 & 14 \\ 1 & 5 & 25 & 12.5 \\ 1 & 10 & 100 & 7 \\ 1 & 10 & 100 & 5 \\ 1 & 15 & 225 & 2.1 \\ 1 & 15 & 225 & 1.8 \\ 1 & 20 & 400 & 6.2 \\ 1 & 20 & 400 & 4.9 \\ 1 & 25 & 625 & 13.2 \\ 1 & 25 & 625 & 14.6 \end{array} \right\}$$

A Polynomial Model

Then \mathbf{b} is given by

```
b = Inverse[Transpose[X].X].Transpose[X].y;
MatrixForm[b]
```

$$\begin{pmatrix} 27.3 \\ -3.313 \\ 0.111 \end{pmatrix}$$

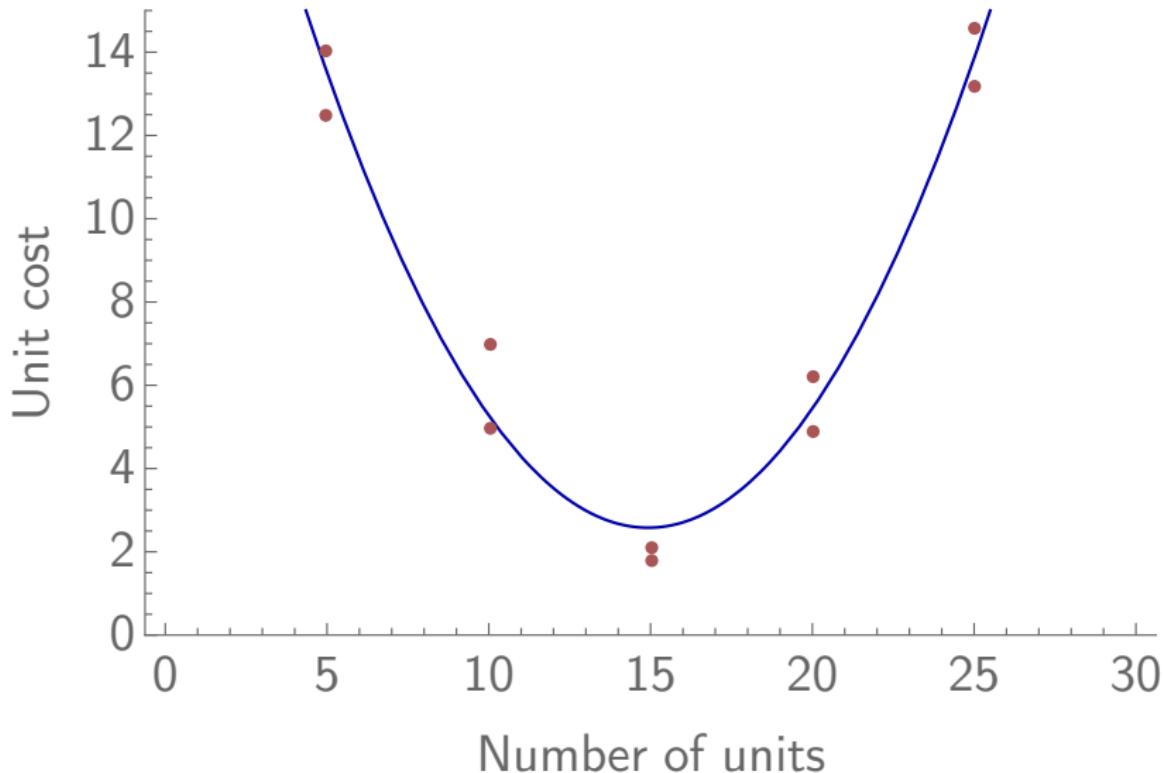
Thus we obtain

$$\hat{\mu}_{Y|x} = 27.3 - 3.313 \cdot x + 0.111 \cdot x^2.$$

The same result can also be found by using **NonLinearModelFit**:

```
model = NonlinearModelFit[data, b0 + b1x + b2x2,
    {b0, b1, b2}, x];
model["BestFit"]
```

$$27.3 - 3.313x + 0.111x^2$$

 A Polynomial Model

A Polynomial Model

We can also use **LinearModelFit** with a given model specification matrix X (called a *design matrix* in Mathematica) and the response vector y :

```
model = LinearModelFit[{X, y}];  
model["BestFit"]
```

```
27.3 #1 - 3.313 #2 + 0.111 #3
```

The output is a “pure function” with three arguments. To obtain the desired expression, we need to insert the appropriate monomials:

```
Evaluate[model["BestFit"]]&[1, x, x2]
```

```
27.3 - 3.313 x + 0.111 x2
```

The Multilinear Model

In the multilinear model, we assume that Y depends on several factors x_1, \dots, x_p ,

$$Y | x = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \dots + \beta_p x_p + E.$$

We take a random sample $(x_{1i}, x_{2i}, \dots, x_{pi}; Y | x_{1i}, x_{2i}, \dots, x_{pi}), i = 1, \dots, n$, writing $Y_i = Y | x_{1i}, x_{2i}, \dots, x_{pi}$ as usual.

We select b_0, \dots, b_p such that for

$$y_i = b_0 + b_1 x_{1i} + \dots + b_p x_{pi} + e_i, \quad i = 1, \dots, n, \quad (26.7)$$

the sum of squares error

$$SS_E = \sum_{i=1}^n e_i^2 = \sum_{i=1}^n (y_i - (b_0 + b_1 x_{1i} + \dots + b_p x_{pi}))^2 \quad (26.8)$$

is minimized.

The Multilinear Model

In fact, the situation is identical to the polynomial model if the model determination matrix X is replaced by

$$X = \begin{pmatrix} 1 & x_{11} & \dots & x_{p1} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & x_{1n} & \dots & x_{pn} \end{pmatrix}.$$

We again have

$$\mathbf{Y} = X\boldsymbol{\beta} + \mathbf{E}.$$

and estimate $\boldsymbol{\beta}$ by minimizing

$$\text{SS}_E = (\mathbf{Y} - X\mathbf{b})^T(\mathbf{Y} - X\mathbf{b}). \quad (26.9)$$

All following calculations remain unchanged and we obtain

$$\mathbf{b} = (X^T X)^{-1} X^T \mathbf{Y}.$$

 A Multilinear Model

26.2. Example. An equation is to be developed from which we can predict the gasoline mileage of an automobile based on its weight and temperature at the time of operation. The model being estimated is

$$\mu_{Y|x} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

These data are available:

Car number	1	2	3	4	5	6	7	8	9	10
Weight in tons (x_1)	1.35	1.90	1.70	1.80	1.30	2.05	1.60	1.80	1.85	1.40
Temp. in $^{\circ}\text{F}$ (x_2)	90	30	80	40	35	45	50	60	65	30
Miles/Gallon (y)	17.9	16.5	16.4	16.8	18.8	15.5	17.5	16.4	15.9	18.3

We enter the data as follows:

```
rowdata :=  
{{1.35, 1.90, 1.70, 1.80, 1.30, 2.05, 1.60, 1.80, 1.85, 1.40},  
 {90, 30, 80, 40, 35, 45, 50, 60, 65, 30},  
 {17.9, 16.5, 16.4, 16.8, 18.8, 15.5, 17.5, 16.4, 15.9, 18.3}}
```

A Multilinear Model

We construct the specification matrix and response vector before obtaining the model parameters:

```
X = Transpose[{Table[1, {i, 1, Length[rowdata[[1]]]}],  
    rowdata[[1]], rowdata[[2]]}];  
y = rowdata[[3]];  
{MatrixForm[X], MatrixForm[y]}
```

$$\left\{ \begin{array}{ccc|c} 1 & 1.35 & 90 & 17.9 \\ 1 & 1.9 & 30 & 16.5 \\ 1 & 1.7 & 80 & 16.4 \\ 1 & 1.8 & 40 & 16.8 \\ 1 & 1.3 & 35 & 18.8 \\ 1 & 2.05 & 45 & 15.5 \\ 1 & 1.6 & 50 & 17.5 \\ 1 & 1.8 & 60 & 16.4 \\ 1 & 1.85 & 65 & 15.9 \\ 1 & 1.4 & 30 & 18.3 \end{array} \right\}$$

 A Multilinear Model

```
b = Inverse[Transpose[X].X].Transpose[X].y; MatrixForm[b]
```

$$\begin{pmatrix} 24.7489 \\ -4.15933 \\ -0.014895 \end{pmatrix}$$

We could also have used **LinearModelFit** based on the model specification matrix,

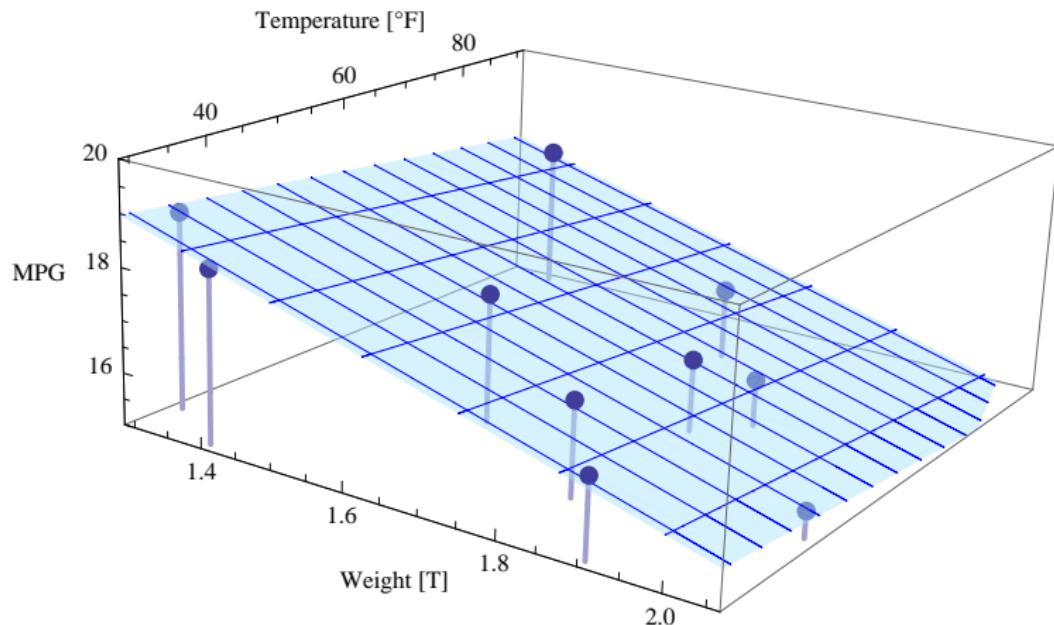
```
model = LinearModelFit[{X, y}];  
Evaluate[model["BestFit"]]] &[1, x1, x2]  
24.7489 - 4.15933 x1 - 0.014895 x2
```

or entered the data directly in the form of a list of triples (x_1, x_2, y) ,

```
data = Transpose[{a1, a2, y}];  
model = LinearModelFit[data, {x1, x2}, {x1, x2}];  
model["BestFit"]  
24.7489 - 4.15933 x1 - 0.014895 x2
```

 A Multilinear Model

The model gives a regression plane:



Error Analysis: Total Variation

Let us now analyze the sources of variation in our models. The total variation is given by

$$SS_T = \sum_{i=1}^n (Y_i - \bar{Y})^2$$

It is convenient to express this in matrix notation: we define the $n \times n$ matrix

$$P := \frac{1}{n} \begin{pmatrix} 1 & 1 & \dots & 1 \\ \vdots & \vdots & & \vdots \\ 1 & 1 & \dots & 1 \end{pmatrix}.$$

Then it is easy to see that

$$PY = \begin{pmatrix} \bar{Y} \\ \vdots \\ \bar{Y} \end{pmatrix}.$$

Error Analysis: The P Projection

This allows us to write

$$SS_T = \langle (\mathbb{1}_n - P)\mathbf{Y}, (\mathbb{1}_n - P)\mathbf{Y} \rangle$$

where $\mathbb{1}_n$ is the $n \times n$ unit matrix. We further remark that

$$P^2 = P, \quad \text{and} \quad P^T = P. \quad (26.10)$$

A matrix with the properties (26.10) is said to be an **orthogonal projection**. We can easily check that (26.10) implies

$$(\mathbb{1}_n - P)^2 = \mathbb{1}_n - P, \quad \text{and} \quad (\mathbb{1}_n - P)^T = \mathbb{1}_n - P.$$

Then

$$SS_T = \langle \mathbf{Y}, (\mathbb{1}_n - P)^T (\mathbb{1}_n - P)\mathbf{Y} \rangle = \langle \mathbf{Y}, (\mathbb{1}_n - P)\mathbf{Y} \rangle \quad (26.11)$$

Such an expression is called a **quadratic form** in \mathbf{Y} .

The Hat Matrix

Recall that both the polynomial and the multilinear models were based on writing the n responses in the form

$$\mathbf{Y} = \mathbf{X}\mathbf{b} + \mathbf{e}$$

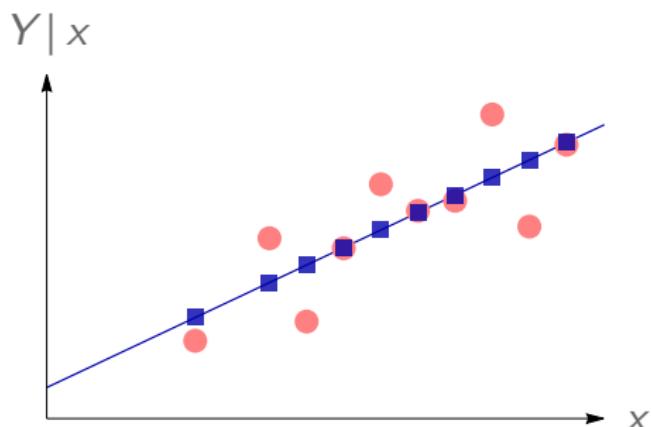
where \mathbf{e} is the least-squares vector of residuals and $\mathbf{Y} = (Y_1, \dots, Y_n)^T$ is the vector of the responses.

● Y_i ■ \hat{Y}_i

The vector

$$\hat{\mathbf{Y}} := \mathbf{X}\mathbf{b}$$

then represents the predicted values of the responses, i.e., the points \hat{Y}_i lying on the regression curve at x_i .



The Hat Matrix

Since $\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$ we may write

$$\hat{\mathbf{Y}} = H\mathbf{Y}, \quad H := \mathbf{X}(\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T,$$

where the $n \times n$ matrix H is called the **hat matrix**. It associates to each measured response \mathbf{Y}_i the predicted response $\hat{\mathbf{Y}}_i$.

Like P , the hat matrix is an orthogonal projection: we can check that

$$H\mathbf{X} = \mathbf{X}, \quad H^T = H, \quad H^2 = H.$$

Therefore, so is $\mathbb{1}_n - H$ and we have

$$(\mathbb{1}_n - H)\mathbf{X} = 0, \quad (\mathbb{1}_n - H)^T = \mathbb{1}_n - H, \quad (\mathbb{1}_n - H)^2 = \mathbb{1}_n - H. \quad (26.12)$$

Error Analysis: Sum of Squares Error

We may therefore write the error sum of squares as

$$\begin{aligned} SS_E &= \langle \mathbf{Y} - X\mathbf{b}, \mathbf{Y} - X\mathbf{b} \rangle \\ &= \langle (\mathbb{1}_n - H)\mathbf{Y}, (\mathbb{1}_n - H)\mathbf{Y} \rangle \\ &= \langle \mathbf{Y}, (\mathbb{1}_n - H)^T(\mathbb{1}_n - H)\mathbf{Y} \rangle \\ &= \langle \mathbf{Y}, (\mathbb{1}_n - H)\mathbf{Y} \rangle. \end{aligned}$$

We may therefore write out a sums of squares decomposition very easily:

$$\begin{aligned} SS_T &= \langle \mathbf{Y}, (\mathbb{1}_n - P)\mathbf{Y} \rangle \\ &= \underbrace{\langle \mathbf{Y}, (\mathbb{1}_n - H)\mathbf{Y} \rangle}_{=:SS_E} + \underbrace{\langle \mathbf{Y}, (H - P)\mathbf{Y} \rangle}_{=:SS_R}. \end{aligned}$$

Fundamental Sum-of-Squares Decomposition

We hence have the decomposition

$$SS_T = SS_R + SS_E \quad (26.13)$$

where

- (i) SS_T represents the total variation of the response variable Y ,
- (ii) SS_R (called the **regression sum of squares**) represents the variation of the response predicted by the regression model and
- (iii) SS_E represents the deviation of the response from the model.

Analogously to (25.2), the **coefficient of multiple determination**,

$$R^2 = \frac{SS_R}{SS_T}$$

gives the proportion of the response variation in Y explained by the model.

Fundamental Sum-of-Squares Error Decomposition

26.3. Remark. It can be shown that

$$SS_R = \langle \mathbf{Y}, (H - P)\mathbf{Y} \rangle = \langle (H - P)\mathbf{Y}, (H - P)\mathbf{Y} \rangle.$$

Then the equation

$$SS_T = SS_R + SS_E \tag{26.14}$$

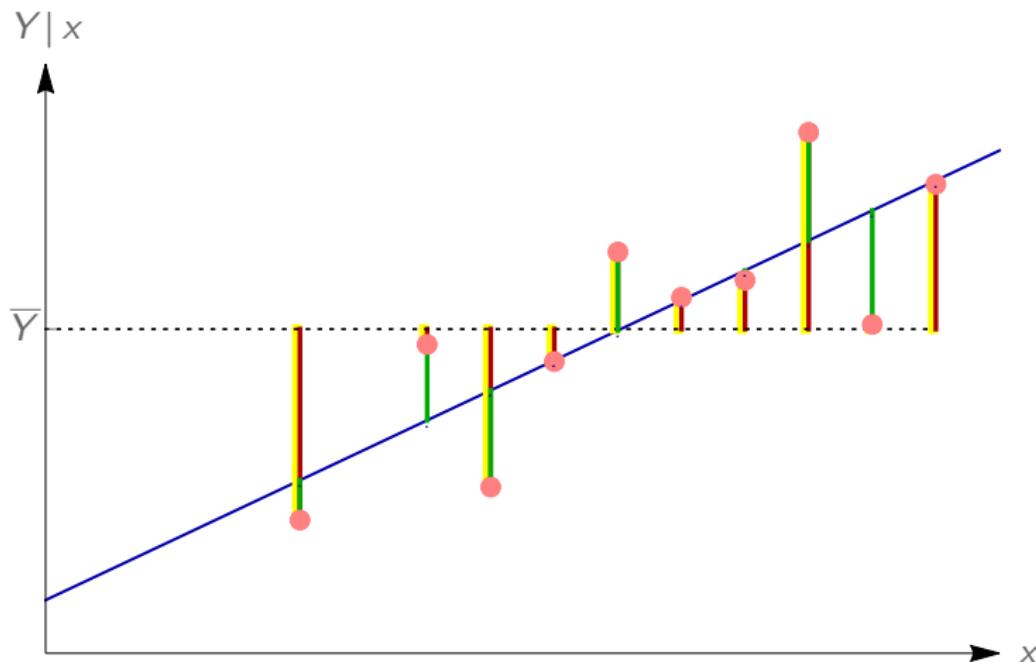
may be expressed as

$$\sum_{i=1}^n (Y_i - \bar{Y})^2 = \sum_{i=1}^n (\hat{Y}_i - \bar{Y})^2 + \sum_{i=1}^n (Y_i - \hat{Y}_i)^2$$

with $\hat{\mathbf{Y}} = \mathbf{X}\mathbf{b}$. Proving this inequality using only elementary algebraic manipulations is a daunting task.

Fundamental Sum-of-Squares Error Decomposition

$$\sum(\text{yellow lengths})^2 = \sum(\text{green lengths})^2 + \sum(\text{red lengths})^2$$



Multiple Linear Regression II: Inferences on the Model

Distribution of the Sum of Squares Error

Model assumptions:

- ▶ $Y | x$ follows a normal distribution with variance σ^2 and mean given by the model.
- ▶ $Y | x$ is independent of $Y | x'$ for $x \neq x'$.

(Here x may be a vector of several different factors or a single factor.)

Goal: Find the distribution of the error sum of squares

$$SS_E = \langle \mathbf{Y}, (\mathbb{1}_n - H)\mathbf{Y} \rangle$$

where $\mathbf{Y} = (Y_1, \dots, Y_n)$ is the response vector and

$$H := X(X^T X)^{-1}X^T$$

is the hat matrix. Here X is the $(p + 1) \times n$ model specification matrix.

Trace of $\mathbb{1}_n - H$

We first need a basic result from linear algebra:

27.1. Lemma. Let $P: \mathbb{R}^n \rightarrow \mathbb{R}^n$ be a projection, i.e., $P^2 = P$. Then the eigenvalues of P may only have values of 0 or 1.

Proof.

Suppose that $Pv = \lambda v$ for some $v \in \mathbb{R}^n$, $v \neq 0$, and $\lambda \in \mathbb{R}$. Then

$$\lambda v = Pv = P^2 v = P(\lambda v) = \lambda(Pv) = \lambda^2 v$$

so $\lambda = \lambda^2$, i.e., $\lambda = 0$ or $\lambda = 1$.



Trace of $\mathbb{1}_n - H$

Recall that the trace of a square $n \times n$ matrix $A = (a_{ij})$ is defined as

$$\text{tr } A := \sum_{i=1}^n a_{ii}.$$

We will use the properties

$$\text{tr}(A + B) = \text{tr } A + \text{tr } B, \quad \text{tr}(AB) = \text{tr}(BA)$$

for square $n \times n$ matrices A, B . Furthermore,

$$\text{tr } A = \text{sum of the eigenvalues of } A.$$

We have

$$\begin{aligned} \text{tr } H &= \text{tr}(X(X^T X)^{-1} X^T) = \text{tr}((X^T X)^{-1} X^T X) \\ &= \text{tr}(\mathbb{1}_{p+1}) = p + 1. \end{aligned}$$

so

$$\text{tr}(\mathbb{1}_n - H) = \text{tr } \mathbb{1}_n - \text{tr } H = n - p - 1$$

Eigenvalues of $\mathbb{1}_n - H$

Since $\mathbb{1}_n - H$ is a projection, the sum of its eigenvalues is also equal to the number of eigenvalues that equal 1.

Hence, $n - p - 1$ eigenvalues of $\mathbb{1}_n - H$ are equal to 1 and $p + 1$ eigenvalues equal 0.

Since $\mathbb{1}_n - H$ is symmetric, we can apply the spectral theorem of linear algebra: there exists a matrix U (whose columns are eigenvectors of $\mathbb{1}_n - H$) such that

$$U^{-1} = U^T$$

and

$$U^T(\mathbb{1}_n - H)U = \begin{pmatrix} \mathbb{1}_{n-p-1} & 0 \\ 0 & 0 \end{pmatrix} =: D_{n-p-1} \quad (27.1)$$

Distribution of the Sum of Squares Error

Recall that in our model the response vector satisfies

$$\mathbf{Y} = \mathbf{X}\boldsymbol{\beta} + \mathbf{E}$$

where \mathbf{E} follows a normal distribution with mean 0 and variance σ^2 .

Since $\mathbb{1}_n - H$ is an orthogonal projection and $(\mathbb{1}_n - H)\mathbf{X} = 0$ (see (26.12)) and we find

$$\begin{aligned} SS_E &= \langle (\mathbb{1}_n - H)\mathbf{Y}, (\mathbb{1}_n - H)\mathbf{Y} \rangle \\ &= \langle (\mathbb{1}_n - H)(\mathbf{X}\boldsymbol{\beta} + \mathbf{E}), (\mathbb{1}_n - H)(\mathbf{X}\boldsymbol{\beta} + \mathbf{E}) \rangle \\ &= \langle (\mathbb{1}_n - H)\mathbf{E}, (\mathbb{1}_n - H)\mathbf{E} \rangle \\ &= \langle \mathbf{E}, (\mathbb{1}_n - H)\mathbf{E} \rangle \end{aligned}$$

Distribution of the Sum of Squares Error

Since each E_j follows an independent normal distribution with mean zero and variance σ^2 , we have

$$\frac{SS_E}{\sigma^2} = \left\langle \frac{\mathbf{E}}{\sigma}, (\mathbb{1}_n - H) \left(\frac{\mathbf{E}}{\sigma} \right) \right\rangle = \langle \mathbf{Z}, (\mathbb{1}_n - H)\mathbf{Z} \rangle$$

where $\mathbf{Z} = (Z_1, \dots, Z_n)^T$ is a vector of i.i.d. standard normal random variables.

We now use the diagonalization (27.1),

$$\begin{aligned} \frac{SS_E}{\sigma^2} &= \langle \mathbf{Z}, U^T D_{n-p-1} U \mathbf{Z} \rangle = \langle U\mathbf{Z}, D_{n-p-1} U\mathbf{Z} \rangle \\ &= \sum_{i=1}^{n-p-1} (U\mathbf{Z})_i^2 \end{aligned}$$

Since each Z_j follows an independent standard normal distribution, so does each component of $U\mathbf{Z}$. We conclude immediately that SS_E follows a chi-squared distribution with $n - p - 1$ degrees of freedom.

Distribution of the Sum of Squares Error

We can apply analogous arguments to SS_R and SS_T . In summary, we have

27.2. Theorem.

- (i) SS_E / σ^2 follows a chi-squared distribution with $n - p - 1$ degrees of freedom.
- (ii) If $\beta = (\beta_0, 0, \dots, 0)$, then SS_R / σ^2 follows a chi-squared distribution with p degrees of freedom.

Furthermore, SS_R and SS_E are independent random variables.

27.3. Corollary. The estimator

$$S^2 := \frac{SS_E}{n - p - 1}$$

is unbiased for σ^2 .

Practical Calculations

27.4. Lemma. The regression sum of squares can be expressed as

$$SS_R = \langle \mathbf{b}, X^T Y \rangle - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2$$

In particular, in the case of the multilinear model,

$$SS_R = b_0 \sum_{i=1}^n Y_i + \sum_{j=1}^p b_j \sum_{i=1}^n x_{ji} Y_i - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2,$$

and in the polynomial model,

$$SS_R = b_0 \sum_{i=1}^n Y_i + \sum_{j=1}^p b_j \sum_{i=1}^n x_i^j Y_i - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2.$$

Estimated Variance and Correlation Coefficient

27.5. Example. In Example 26.2 we obtained the regression equation

$$\hat{\mu}_{Y|x_1, x_2} = 24.75 - 4.16x_1 - 0.015x_2.$$

for the mean gas mileage of cars as a function of weight x_1 and motor temperature x_2 . We now want to find R^2 for our model.

It is convenient to write

$$SS_R = \langle B, X^T Y \rangle - \frac{1}{n} \left(\sum_{i=1}^n Y_i \right)^2.$$

We first calculate

$$X^T y = \begin{pmatrix} \sum y_i \\ \sum x_{1i}y_i \\ \sum x_{2i}y_i \end{pmatrix} = \begin{pmatrix} 170.00 \\ 282.405 \\ 8887.00 \end{pmatrix}, \quad \sum_{i=1}^n y_i^2 = 2900.46.$$

Estimated Variance and Coefficient of Determination

These values give us

$$SS_T = \sum_{i=1}^n y_i^2 - \frac{1}{n} \left(\sum_{i=1}^n y_i \right)^2 = 10.46,$$

$$SS_R = \left\langle \begin{pmatrix} 170.00 \\ 282.405 \\ 8887.00 \end{pmatrix}, \begin{pmatrix} 24.75 \\ -4.16 \\ -0.015 \end{pmatrix} \right\rangle - \frac{170.00^2}{25} = 10.32.$$

Hence,

$$R^2 = \frac{10.32}{10.46} = 0.9866.$$

We also note that $SS_E = SS_{tot} - SS_R = 10.46 - 10.32 = 0.14$ and the estimated variance is

$$\hat{\sigma}^2 = s^2 = \frac{SS_E}{n - p - 1} = \frac{0.14}{10 - 2 - 1} = 0.02.$$



Estimated Variance and Coefficient of Determination

We can extract the estimated variance and R^2 directly from the model:

```
model = LinearModelFit[data, {x1, x2}, {x1, x2}];  
model["EstimatedVariance"]
```

0.02005

```
model["RSquared"]
```

0.986582

F-Test for Significance of Regression

Since SS_R measures the variability associated with the model and SS_E measures “random variation”, we will find the regression significant if SS_R is much larger than SS_E . The basis for the test is Theorem 27.2.

27.6. F-Test for Significance of Regression. Let x_1, \dots, x_p be the predictor variables in a multilinear model (26.1) for Y . Then

$$H_0: \beta_1 = \beta_2 = \dots = \beta_p = 0,$$

is rejected at significance level α if the test statistic

$$F_{p,n-p-1} = \frac{SS_R / p}{SS_E / (n - p - 1)} = \frac{SS_R / p}{S^2} \quad (27.2)$$

satisfies $F_{p,n-p-1} > f_{\alpha,p,n-p-1}$.

Significance of Regression

We remark that

$$\begin{aligned} F_{p,n-p-1} &= \frac{n-p-1}{p} \frac{SS_R / SS_T}{SS_E / SS_T} = \frac{n-p-1}{p} \frac{SS_R / S_{yy}}{(SS_T - SS_R) / SS_T} \\ &= \frac{n-p-1}{p} \frac{R^2}{1-R^2} \end{aligned}$$

so the value of R^2 alone can be used to test for significance of regression.

27.7. Example. In Example 27.5, we obtained $R^2 = 0.986$. Since $n = 10$ and $p = 2$ the value of the test statistic for significance of regression is

$$\frac{n-p-1}{p} \frac{R^2}{1-R^2} = \frac{7}{2} \frac{0.986}{0.014} = 243.05.$$

The 95% point of the $F_{2,7}$ -distribution is 4.74, so we can reject H_0 with $P < 0.05$. There is evidence that the regression is significant.

Expectation for Random Vectors

Goal: Derive distribution of the model parameters β .

Recall: Let $Y = (Y_1, \dots, Y_n)^T$ be a random vector. Then

$$\mathbb{E}[Y] = \begin{pmatrix} \mathbb{E}[Y_1] \\ \vdots \\ \mathbb{E}[Y_n] \end{pmatrix}.$$

For random vectors Y, Z and a constant $m \times n$ matrix C :

- (i) $\mathbb{E}[C] = C$,
- (ii) $\mathbb{E}[CY] = C\mathbb{E}[Y]$,
- (iii) $\mathbb{E}[Y + Z] = \mathbb{E}[Y] + \mathbb{E}[Z]$.

Expectation of the Least-Squares Estimators

We can calculate directly that the expectation of the response vector is

$$\mathbb{E}[\mathbf{Y}] = \mathbb{E}[X\beta + \mathbf{E}] = \mathbb{E}[X\beta] + \mathbb{E}[\mathbf{E}] = X\beta.$$

Then

$$\begin{aligned}\mathbb{E}[\mathbf{b}] &= \mathbb{E}[(X^T X)^{-1} X^T \mathbf{Y}] = (X^T X)^{-1} X^T \mathbb{E}[\mathbf{Y}] \\ &= (X^T X)^{-1} X^T X\beta \\ &= \beta.\end{aligned}$$

It follows that $\hat{\beta} = \mathbf{b}$ is an unbiased estimator for β .

Variance for Random Vectors

Recall: Let $Y = (Y_1, \dots, Y_n)^T$ be a random vector. Then

$$\text{Var}[Y] = \begin{pmatrix} \text{Var}[Y_1] & \text{Cov}[Y_1, Y_2] & \cdots & \text{Cov}[Y_1, Y_n] \\ \text{Cov}[Y_1, Y_2] & \text{Var}[Y_2] & \ddots & \vdots \\ \vdots & \ddots & \ddots & \text{Cov}[Y_{n-1}, Y_n] \\ \text{Cov}[Y_1, Y_n] & \cdots & \text{Cov}[Y_{n-1}, Y_n] & \text{Var}[Y_n] \end{pmatrix}$$

and

$$\text{Var}[CY] = C \text{Var}[Y] C^T,$$

where C is a constant $m \times n$ matrix.

Variance of the Least-Squares Estimators

In our case, a random sample $(x_1, Y_1), \dots, (x_n, Y_n)$ is given where the Y_i are independent and all Y_i have the same variance σ^2 .

Therefore,

$$\text{Var}[\mathbf{Y}] = \sigma^2 \mathbb{1}_n.$$

We then have

$$\begin{aligned}\text{Var}[\mathbf{b}] &= \text{Var}[(X^T X)^{-1} X^T \mathbf{Y}] \\&= (X^T X)^{-1} X^T \text{Var}[\mathbf{Y}] ((X^T X)^{-1} X^T)^T \\&= \sigma^2 (X^T X)^{-1} X^T ((X^T X)^{-1} X^T)^T \\&= \sigma^2 (X^T X)^{-1}\end{aligned}$$

Variance of the Least-Squares Estimators

Let us write

$$X^T X = \begin{pmatrix} \xi_{00} & * & \cdots & \cdots & * \\ * & \xi_{11} & \ddots & & * \\ \vdots & \ddots & \ddots & \ddots & \vdots \\ \vdots & & \ddots & \ddots & * \\ * & \cdots & \cdots & * & \xi_{pp} \end{pmatrix}$$

where the starred values are uninteresting for us.

Hence,

$$\text{Var}[B_i] = \xi_{ii}\sigma^2, \quad i = 0, \dots, p,$$

Note that the estimators B_0, \dots, B_p are not independent of each other, but we will not investigate their covariance here.

Distribution of the Least-Squares Estimators

Since

$$\mathbf{b} = (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

and the components of \mathbf{Y} follow normal distributions, each b_i is a linear combination of independent normal distributions. Hence, each b_i must itself follow a normal distribution.

We have therefore proved the following result:

27.8. Theorem. The random vector \mathbf{b} follows a normal distribution with mean $\boldsymbol{\beta}$ and variance-covariance matrix $\sigma^2(\mathbf{X}^T \mathbf{X})^{-1}$.

It is also possible to prove:

27.9. Theorem. The statistic $(n - p - 1)S^2/\sigma^2 = SS_E/\sigma^2$ is independent of \mathbf{b} .

Confidence Intervals for the Model Parameters

The variables

$$Z = \frac{b_j - \beta_j}{\sigma \sqrt{\xi_{jj}}}, \quad j = 0, \dots, p,$$

are standard normal. Thus, for $j = 0, \dots, p$,

$$\frac{(b_j - \beta_j)/(\sigma \sqrt{\xi_{jj}})}{\sqrt{(n-p-1)S^2/\sigma^2/(n-p-1)}} = \frac{\hat{\beta}_j - \beta_j}{S \sqrt{\xi_{jj}}}, \quad (27.3)$$

follows a T -distribution with $n - p - 1$ degrees of freedom.

We immediately obtain the following $100(1 - \alpha)\%$ confidence intervals for the model parameters:

$$\beta_j = b_j \pm t_{\alpha/2, n-p-1} S \sqrt{\xi_{jj}}, \quad j = 0, \dots, p.$$



Confidence Intervals for the Model Parameters

27.10. Example. Continuing from Example 27.5, we have $s^2 = 0.02005$, so the variance-covariance matrix is

```
MatrixForm[0.02005 Inverse[Transpose[X].X]]
```

$$\begin{pmatrix} 0.121719 & -0.060669 & -0.000344635 \\ -0.060669 & 0.0348589 & 0.0000434344 \\ -0.000344635 & 0.0000434344 & 5.17872 \times 10^{-6} \end{pmatrix}$$

We can also obtain the matrix directly from the model:

```
MatrixForm[model["CovarianceMatrix"]]
```

$$\begin{pmatrix} 0.121719 & -0.0606689 & -0.000344635 \\ -0.0606689 & 0.0348589 & 0.0000434344 \\ -0.000344635 & 0.0000434344 & 5.17871 \times 10^{-6} \end{pmatrix}$$

Confidence Intervals for the Model Parameters

Reading off from the diagonal, we find the estimated variances of the estimators:

$$\widehat{\text{Var}[B_0]} = s^2 \xi_{00} = 0.1217,$$

$$\widehat{\text{Var}[B_1]} = s^2 \xi_{11} = 0.03485,$$

$$\widehat{\text{Var}[B_2]} = s^2 \xi_{22} = 5.178 \cdot 10^{-6}.$$

We hence have the following 95% confidence intervals:

$$\begin{aligned}\beta_0 &= b_0 \pm t_{0.025, 7} \sqrt{s^2 \xi_{00}} = 24.75 \pm 2.365 \sqrt{0.1217} \\ &= 24.75 \pm 0.825\end{aligned}$$

$$\begin{aligned}\beta_1 &= b_1 \pm t_{0.025, 7} \sqrt{s^2 \xi_{11}} = -4.16 \pm 2.365 \sqrt{0.03485} \\ &= -4.16 \pm 0.44\end{aligned}$$

$$\begin{aligned}\beta_2 &= b_2 \pm t_{0.025, 7} \sqrt{s^2 \xi_{22}} = -0.15 \pm 2.365 \sqrt{5.178 \cdot 10^{-6}} \\ &= -0.15 \pm 0.0054\end{aligned}$$



Confidence Intervals for the Model Parameters

Mathematica can directly give the confidence intervals and the standard deviations of the estimators (the square roots of the diagonal elements of the variance-covariance matrix).

```
model["ParameterConfidenceIntervalTable"]
```

	Estimate	Standard Error	Confidence Interval
1	24.7489	0.348882	{23.9239, 25.5738}
x_1	-4.15933	0.186705	{-4.60082, -3.71785}
x_2	-0.014895	0.00227568	{-0.0202761, -0.00951389}

Confidence Intervals for the Estimated Mean

Let us write

$$\mathbf{x}_0 = \begin{pmatrix} 1 \\ x_{10} \\ \vdots \\ x_{p0} \end{pmatrix} \quad \text{or} \quad \mathbf{x}_0 = \begin{pmatrix} 1 \\ x \\ \vdots \\ x^p \end{pmatrix}$$

depending on whether we are considering a multilinear or a polynomial model. Of course, any combination of the two may be considered analogously.

Our goal is to make inferences on the estimated mean at \mathbf{x}_0 . We write

$$\hat{\mu}_{Y|\mathbf{x}_0} = \mathbf{x}_0^T \mathbf{b} = \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{X}^T \mathbf{Y}$$

We see that $\hat{\mu}_{Y|\mathbf{x}_0}$ is a linear combination of the independent and normally distributed Y_i and therefore follows a normal distribution.

Confidence Intervals for the Estimated Mean

Furthermore,

$$E[\hat{\mu}_{Y|x_0}] = E[x_0^T b] = x_0^T E[b] = x_0^T \beta = \mu_{Y|x_{10}, \dots, x_p}$$

and

$$\text{Var}[\hat{\mu}_{Y|x_0}] = \text{Var}[x_0^T b] = x_0^T \text{Var}[b] x_0 = \sigma^2 x_0^T (X^T X)^{-1} x_0.$$

It follows that

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{\sigma \sqrt{x_0^T (X^T X)^{-1} x_0}}$$

is standard normal and, after dividing by $\sqrt{(n-p-1)S^2/\sigma^2}/\sqrt{n-p-1}$
that

$$\frac{\hat{\mu}_{Y|x_0} - \mu_{Y|x_0}}{S \sqrt{x_0^T (X^T X)^{-1} x_0}} \quad (27.4)$$

follows a T distribution with $n-p-1$ degrees of freedom.

Confidence Intervals for the Estimated Mean

We thus have the following $100(1 - \alpha)\%$ confidence interval for $\mu_{Y|x_0}$:

$$\mu_{Y|x_0} = \hat{\mu}_{Y|x_0} \pm t_{\alpha/2, n-p-1} S \sqrt{x_0^T (X^T X)^{-1} x_0}$$

27.11. Example. Following on from Example 27.5, the estimate for the average gasoline mileage for a car weighing 1.5 tons being operated at 70° F is

$$\hat{\mu}_{Y|1.5,70} = 24.75 - 4.16 \cdot 1.5 - 0.14897 \cdot 70 = 17.47.$$

We want to find a 95% confidence interval for this mean. The vector x_0 is given by

$$x_0 = \begin{pmatrix} 1 \\ 1.5 \\ 70 \end{pmatrix}.$$

Prediction Intervals

Then

$$\mu_{Y|1.5,70} = 17.47 \pm 2.365 \cdot S \sqrt{\mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0} = 17.47 \pm 0.16.$$

This agrees with Mathematica's built-in functionality:

```
model["MeanPredictionBands"] /. {x1 → 1.5, x2 → 70}  
{17.3105, 17.6239}
```

As in the previous section, we can obtain a similar $100(1 - \alpha)\%$ ***prediction interval*** for the value of $Y | x_{10}, \dots, x_{p0}$,

$$Y | \mathbf{x}_0 = \hat{\mu}_{Y|\mathbf{x}_0} \pm t_{\alpha/2, n-p-1} S \sqrt{1 + \mathbf{x}_0^T (\mathbf{X}^T \mathbf{X})^{-1} \mathbf{x}_0}.$$

We omit the (completely analogous) details.

Hypothesis Testing on the Model Parameters

Based on the T -distributions of (27.3) and (27.4) we can of course perform tests on the model parameters β and the predicted mean $\hat{\mu}_{Y|x}$.

Since such tests should be routine by now, we omit the details. However, a special case is of interest:

27.12. T -Test for Model Sufficiency. Suppose that a regression model using the parameters β_0, \dots, β_p is fitted to Y . Then for any $j = 0, \dots, p$

$$H_0: \beta_j = 0$$

is rejected at significance level α if the test statistic

$$T_{n-p-1} = \frac{b_j}{S\sqrt{\xi_{jj}}}.$$

satisfies $|T_{n-p-1}| > t_{\alpha/2, n-p-1}$.

T-Test for the Model Parameters

If we are able to reject H_0 , there is evidence that the predictor is needed for the model.

If we fail to reject H_0 , there is no evidence that the predictor is needed and we may proceed to fit a model without this predictor.

27.13. Example. Suppose we are given the data

x	5	7.5	10	12.5	15	17.5	20
y	1	2.2	4.9	5.3	8.2	10.7	13.2

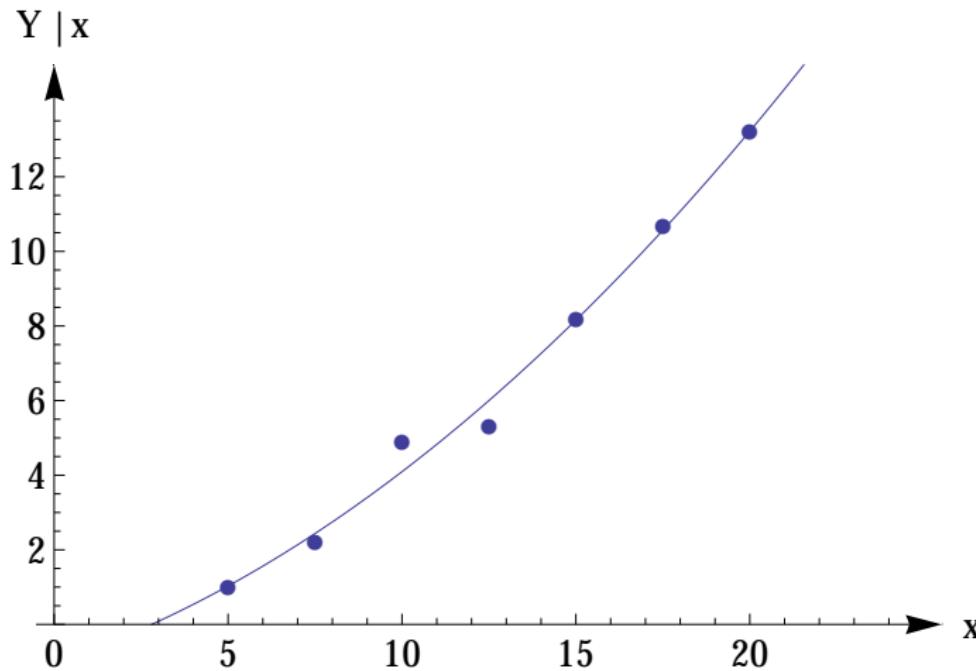
We would like to find a quadratic model for the data:

```
Data = {{5, 1}, {7.5, 2.2}, {10, 4.9`}, {12.5, 5.3`},  
       {15, 8.2`}, {17.5, 10.7}, {20, 13.2`}};  
model = NonlinearModelFit[Data, b0 + b1 x + b2 x^2 {b0, b1, b2}, x];  
model["BestFit"]
```

$$-1.03571 + 0.312857 x + 0.02 x^2$$

T-Test for the Model Parameters

The data and the model curve is plotted below.



T-Test for the Model Parameters

We can find confidence intervals for all model parameters:

```
model["ParameterConfidenceIntervalTable",  
       ConfidenceLevel → 0.975]
```

	Estimate	Standard Error	Confidence Interval
b_0	-1.03571	1.3838	{-5.87265, 3.80122}
b_1	0.312857	0.244475	{-0.541682, 1.1674}
b_2	0.02	0.00963554	{-0.0136801, 0.0536801}

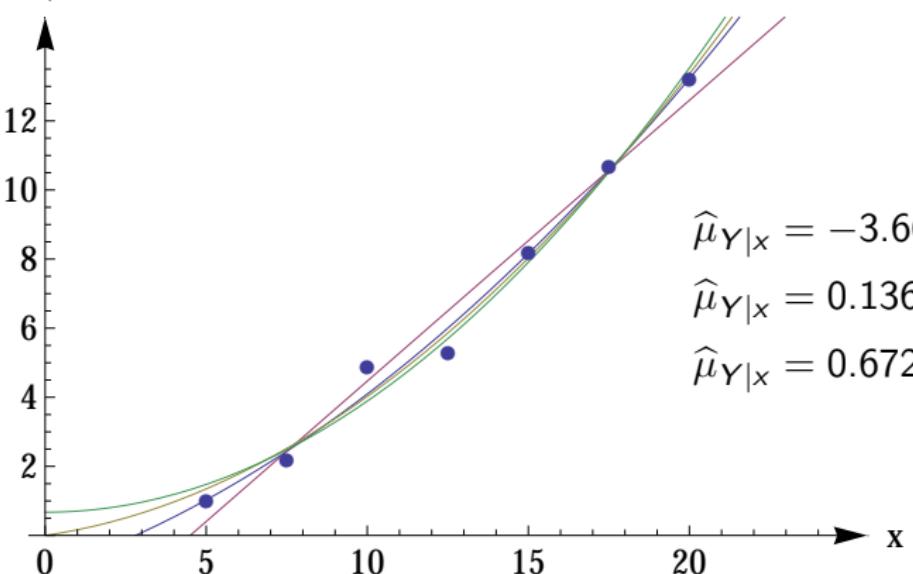
Based on these 95% confidence intervals, we can not reject $H_0: \beta_j = 0$ for any $j = 0, 1, 2$. This means that there is no evidence that any single β_j is non-zero.

However, not all coefficients will be zero. The regression is clearly significant (as can be seen by conducting a test for significance of regression; see Example 27.7).

T-Test for the Model Parameters

We can eliminate any one of the there predictors simply be deleting the corresponding column from the model specification matrix X . This yields the alternative models

Y | x



$$\hat{\mu}_{Y|x} = -3.66071 + 0.812857x,$$

$$\hat{\mu}_{Y|x} = 0.136315x + 0.0265909x^2,$$

$$\hat{\mu}_{Y|x} = 0.67285 + 0.0321498x^2.$$

General Test for Model Sufficiency

It is of course not clear which of these three models is best; this is a question we will return at a later point.

The T -test 27.12 can be used to determine whether a single predictor may be eliminated from the model. It is often practical, however, to compare a general subset of predictors with a full model of $p + 1$ predictor variables,

$$\mu_{Y|x_1, \dots, x_p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_p x_p. \quad (27.5)$$

After possibly renumbering the variables we compare with a **reduced model** of $m + 1 < p + 1$ predictor variables

$$\mu_{Y|x_1, \dots, x_m} = \tilde{\beta}_0 + \tilde{\beta}_1 x_1 + \cdots + \tilde{\beta}_m x_m. \quad (27.6)$$

We define the sums of squares errors for the two models by

$SS_{E;\text{full}}$ = sum of squares error SS_E for full model,

$SS_{E;\text{reduced}}$ = sum of squares error SS_E for reduced model.

Partial F -Test for Model Sufficiency

We will base our test on the principle that there is evidence that the full model is needed if $SS_{E;\text{full}} \ll SS_{E;\text{reduced}}$.

27.14. Partial F -Test for Model Sufficiency. Let x_1, \dots, x_p be possible predictor variables for Y and (27.5) and (27.6) the full and reduced models, respectively. Then

H_0 : the reduced model is sufficient

is rejected at significance level α if the test statistic

$$F_{p-m, n-p-1} = \frac{n-p-1}{p-m} \frac{SS_{E;\text{reduced}} - SS_{E;\text{full}}}{SS_{E;\text{full}}} \quad (27.7)$$

satisfies $F_{p-m, n-p-1} > f_{\alpha, p-m, n-p-1}$.

Partial F -Test for Model Sufficiency

27.15. Example. In the context of Example 27.13 we can compare the linear and quadratic models

$$\hat{\mu}_{Y|x;\text{full}} = -1.03571 + 0.312857x + 0.02x^2, \quad SS_{E;\text{full}} = 1.21857,$$

$$\hat{\mu}_{Y|x;\text{reduced}} = -3.66071 + 0.812857x, \quad SS_{E;\text{reduced}} = 2.53107.$$

Here, $n = 7$, $p = 2$, $m = 1$, so

$$F_{p-m, n-p-1} = \frac{n-p-1}{p-m} \frac{SS_{E;\text{reduced}} - SS_{E;\text{full}}}{SS_{E;\text{full}}} = 4.30832.$$

The critical point $f_{0.05,1,4} = 7.71$, so we can not reject H_0 at the 5% level of significance. There is no evidence that the full model is needed.

Partial F -Test for Model Sufficiency

27.16. Example. Continuing with Example 27.13 we can also compare the general quadratic model with a square monomial model:

$$\hat{\mu}_{Y|x;\text{full}} = -1.03571 + 0.312857x + 0.02x^2, \quad SS_{E;\text{full}} = 1.21857,$$

$$\hat{\mu}_{Y|x;\text{reduced}} = 0.0346414x^2, \quad SS_{E;\text{reduced}} = 1.83967.$$

Here, $n = 7$, $p = 2$, $m = 0$, so

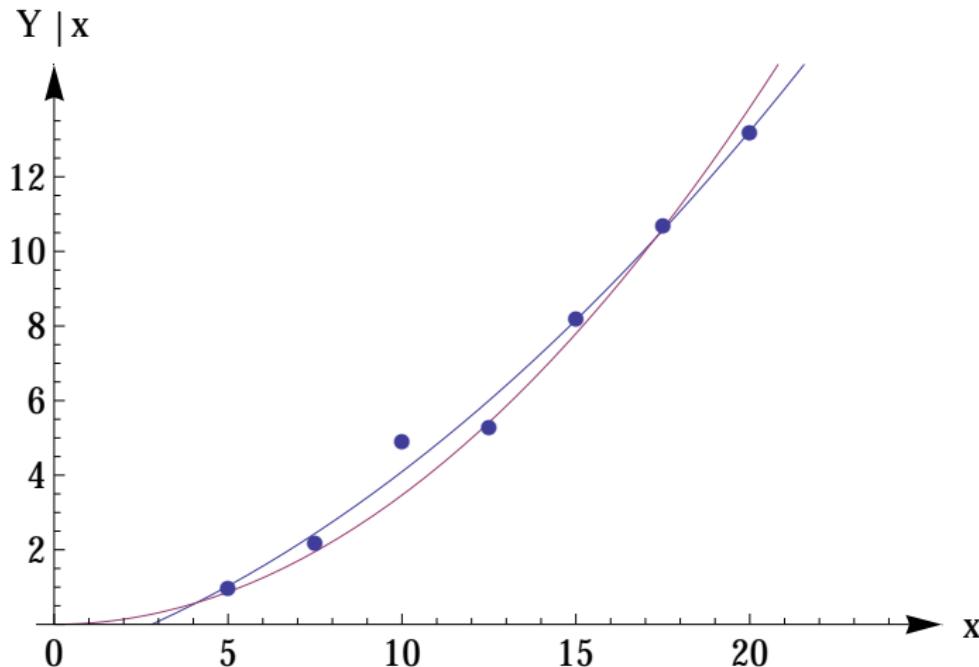
$$F_{p-m, n-p-1} = \frac{n-p-1}{p-m} \frac{SS_{E;\text{reduced}} - SS_{E;\text{full}}}{SS_{E;\text{full}}} = 1.01939.$$

The critical point $f_{0.05,2,4} = 6.94$, so we can not reject H_0 at the 5% level of significance. There is no evidence that the full model is needed.

Comparing the sum of squares errors with the previous example, we can furthermore conclude that a square monomial model gives a better fit than the linear model.

Partial F -Test for Model Sufficiency

The graph below shows the quadratic and the square monomial models.



T-Test and Partial *F*-Test for Single Predictors

While the *T*-test can be used to determine whether a single predictor is necessary for a given model, the *F*-test can be applied to an arbitrary subset of predictors.

The question arises whether there is a difference between the two tests when considering a single predictor, i.e., whether the *F*-test applied to a single variable (as in Example 27.15) always yields the same result as the *T*-test.

It is possible to prove that, indeed, the *T*-test for a single variable is equivalent to a partial *F*-Test when applied to a reduced model lacking only that single variable.

Interpretations of the Partial F -Test

Furthermore, since $SS_T = SS_R + SS_E$, the test statistic (27.7) can be re-written as

$$F_{p-m, n-p-1} = \frac{n-p-1}{p-m} \frac{SS_{E;\text{reduced}} - SS_{E;\text{full}}}{SS_{E;\text{full}}} \\ \frac{n-p-1}{p-m} \frac{SS_{R;\text{full}} - SS_{R;\text{reduced}}}{SS_{E;\text{full}}}.$$

This shows that the F -test for significance of regression based on the statistic (27.2),

$$F_{p, n-p-1} = \frac{n-p-1}{p} \frac{SS_R}{SS_E}$$

may be regarded as a partial F -test where the reduced model contains no regressors.

Moreover, the partial F -test can be formulated in terms of the determination coefficients R^2 for the full and reduced models.

Multiple Linear Regression III: Finding the Right Model

Qualitative Predictors

Problem: Include **categorical predictors** in a regression: brand, type, gender, etc.

Suppose our data is of two different "types", Type A and Type B.

We introduce a parameter (**indicator variable**)

$$X = \begin{cases} 1, & \text{predictor is of type } A, \\ 0, & \text{predictor is of type } B. \end{cases}$$

This indicator variable can be included in regression models, as shown in the following example.

Example: Indicator Variable for the Intercept

28.1. Example. Consider the previously discussed Example 24.1:

Response: solvent evaporation in spray paint Y

Predictors:

- ▶ Humidity x_1
- ▶ Brand of spray paint x_2

Assumption/Model: humidity has the same systematic effect, but the paint may be generally more resistant depending on the brand:

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

where

$$x_2 = \begin{cases} 1, & \text{brand } A \text{ used,} \\ 0, & \text{brand } B \text{ used.} \end{cases}$$

Example: Indicator Variable for the Intercept

Brand A used (x_2)	x_1	y	Brand B used (x_2)	x_1	y
1	35.3	11.2	0	39.1	6.7
1	29.6	11.0	0	46.8	7.7
1	31.0	12.6	0	48.5	6.8
1	58.0	8.3	0	59.3	7.0
1	62.0	10.1	0	70.0	5.2
1	72.1	9.6	0	70.0	4.0
1	74.0	6.1	0	74.4	5.7
1	77.0	8.7	0	72.1	4.9
1	71.1	8.1	0	58.1	5.5
1	57.0	9.0	0	44.6	6.1
1	46.4	8.2	0	33.4	7.5
1	29.6	13.0	0	28.6	8.0
1	28.0	11.7			

x_1 is the observed relative humidity (in %), and y is the observed solvent evaporation (in %).

Example: Indicator Variable for the Intercept

```
humidity = {35.3, 29.6, 31.0, 58.0, 62.0, 72.1, 74.0, 77.0, 71.1, 57.0, 46.4, 29.6,
           28.0, 39.1, 46.8, 48.5, 59.3, 70.0, 70.0, 74.4, 72.1, 58.1, 44.6, 33.4, 28.6};
n = Length[humidity];
X = Transpose[{Table[1, {i, n}], humidity, Join[Table[1, {i, 13}], Table[0, {i, 12}]]}];
MatrixForm[X]
```

1	35.3	1
1	29.6	1
1	31.	1
1	58.	1
1	62.	1
1	72.1	1
1	74.	1
1	77.	1
1	71.1	1
1	57.	1
1	46.4	1
1	29.6	1
1	28.	1
1	39.1	0
1	46.8	0
1	48.5	0
1	59.3	0
1	70.	0
1	70.	0
1	74.4	0
1	72.1	0
1	58.1	0
1	44.6	0
1	33.4	0
1	28.6	0

Example: Indicator Variable for the Intercept

From

```
MatrixForm[Inverse[Transpose[X].X]]
```

$$\begin{pmatrix} 0.488429 & -0.00753783 & -0.0993029 \\ -0.00753783 & 0.00014026 & 0.000297154 \\ -0.0993029 & 0.000297154 & 0.160886 \end{pmatrix}$$

```
y = {11.2, 11.0, 12.6, 8.3, 10.1, 9.6, 6.1, 8.7, 8.1, 9.0,  
     8.2, 13.0, 11.7, 6.7, 7.7, 6.8, 7.0, 5.2, 4.0, 5.7, 4.9,  
     5.5, 6.1, 7.5, 8.0};
```

```
b = Inverse[Transpose[X].X].Transpose[X].y;
```

```
MatrixForm[b]
```

$$\begin{pmatrix} 10.398 \\ -0.0770288 \\ 3.39386 \end{pmatrix}$$

we obtain the regression parameters

$$b_0 = 10.3979, \quad b_1 = -0.0770, \quad b_2 = 3.3938.$$

Example: Indicator Variable for the Intercept

The estimated model is

$$\hat{\mu}_{Y|x_1,x_2} = 10.3979 - 0.770x_1 + 3.3938x_2$$

so when paint A is used, the model is

$$\hat{\mu}_{Y|x_1,1} = 13.7917 - 0.770x_1,$$

while the model for paint B is

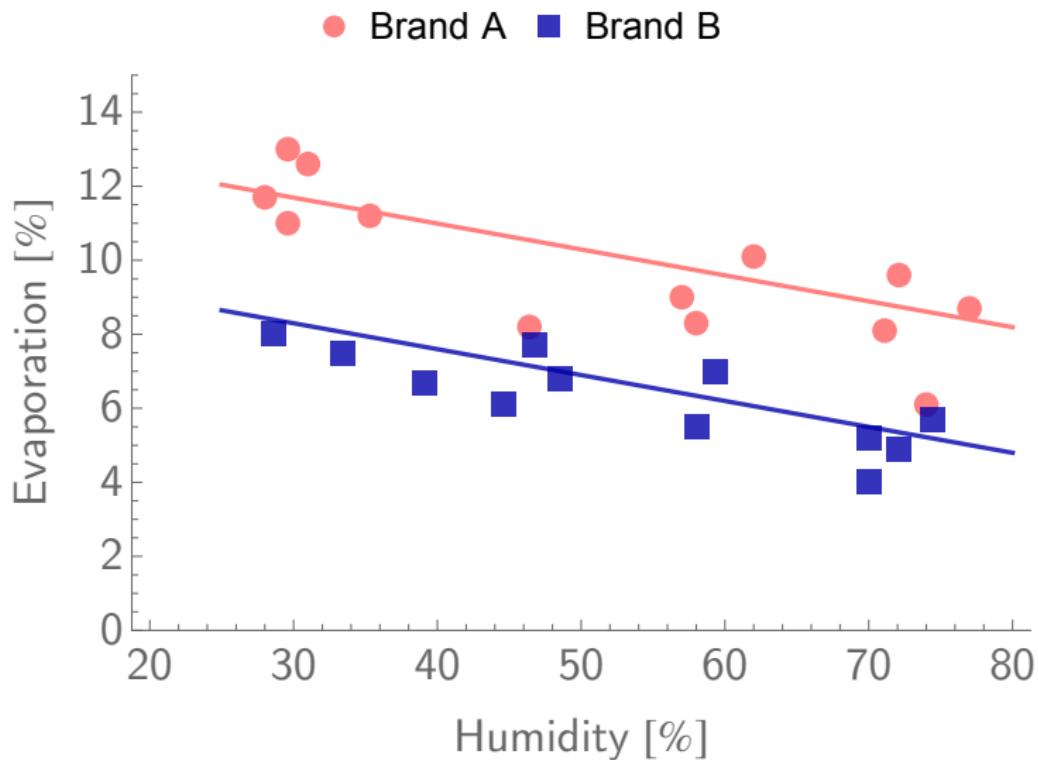
$$\hat{\mu}_{Y|x_1,0} = 10.3979 - 0.770x_1.$$

We could check as usual whether there is evidence to reject

$$H_0: \beta_2 = 0,$$

i.e., whether the brand of paint truly matters.

Example: Indicator Variable for the Intercept



Motivation for Indicator Variables

Why are we doing this?

We could also simply do two separate regressions, one for each brand of paint.

Advantages:

- ▶ Greater overall sample size gives more degrees of freedom, so confidence intervals are tighter and hypothesis tests are more powerful.
- ▶ The brand may be considered as one predictor among many possible predictors, both continuous variables and qualitative variables. It allows for a systematic model selection by comparing full and reduced models.

Indicator Variables for Several Predictors

We can use several indicator variables if there is more than one category or type.

For example, in order to test three brands of paint, we employ a model

$$\mu_Y|x_1, x_2, x_3 = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

where

$$(x_2, x_3) = \begin{cases} (0, 0) & \text{type } A \text{ used,} \\ (1, 0) & \text{type } B \text{ used,} \\ (0, 1) & \text{type } C \text{ used.} \end{cases}$$

The number of possibilities for a qualitative variable are called **levels**. To model ℓ levels, we need $\ell - 1$ indicator variables.

Indicator Variables for Slope and Intercept

In our example we have assumed that the slope of the regression line will be identical. If we do not suppose this to be the case, we can use our indicator variables to also contribute to the slope. In the case of one indicator variable x_2 with two levels we use

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_1 x_2.$$

If $x_2 = 1$, the model is

$$\mu_{Y|x_1,1} = \beta_0 + \beta_2 + (\beta_1 + \beta_3)x_1,$$

while for $x_2 = 0$, the model becomes

$$\mu_{Y|x_1,0} = \beta_0 + \beta_1 x_1.$$

To test for equality of slopes, we test $H_0: \beta_3 = 0$.

The Model Selection Problem

Problem: Select the “right” model:

- ▶ In polynomial regression, the degree of the polynomial must be decided upon;
- ▶ In multiple linear regression, the simplest model through use of the smallest number of predictors must be found.

The basic problem is to find a model that gives a “good fit.”

Naive approach: Maximize R^2 .

Extreme result:

- ▶ In a multilinear model, include every possible predictor.
- ▶ In a polynomial model, let $p = n - 1$ and interpolate the data.

Clearly, this is nonsense. But why?

Model Selection

We don't create a model for its own sake, but because we want to use it!

For example, a confidence interval for $\mu_{Y|x_0}$ is given by

$$\mu_{Y|x_0} = \hat{\mu}_{Y|x_0} \pm t_{\alpha/2, n-p-1} S \sqrt{x_0^T (X^T X)^{-1} x_0}$$

where

$$S^2 = \frac{SS_E}{n - p - 1}.$$

By increasing p , we

- ▶ decrease SS_E and
- ▶ decrease $n - p - 1$.

The second effect is bad for $t_{\alpha/2, n-p-1}$ but can be catastrophic for S^2 .

If p is too large, the model becomes useless.

Model Selection Algorithms

Therefore, we want to increase p only until a further decrease of SS_E is outweighed by the decrease on $n - p - 1$.

More generally, we want to achieve a small SS_E using the smallest possible number of predictors.

One approach is to use a ***model selection algorithm***. A subset of possible models is compared until an “optimal” model is obtained.

We now look at three typical algorithms:

Forward Selection: Variables are added to the model one at a time until the addition of another variable does not significantly improve the model. That is, variables are added until we are unable to reject the reduced model.

Forward Selection Method

28.2. Example. Assume that we have available three possible predictor variables X_1 , X_2 and X_3 . Suppose that our final model via forward selection contains only the variables X_3 and X_1 and that they entered the model in the order stated. These are the steps taken:

1. The three single-variable models

$$\mu_{Y|x_1} = \beta_0 + \beta_1 x_1, \quad \mu_{Y|x_2} = \beta_0 + \beta_2 x_2, \quad \mu_{Y|x_3} = \beta_0 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each.

The one with the highest R^2 is chosen and compared to the reduced model $\mu_Y = \beta_0$. In this case it is the third model and we test

$$H_0: \beta_3 = 0.$$

We find that H_0 is rejected. Our model now includes X_3 .

Forward Selection Method

2. The two two-variable models

$$\mu_{Y|x_3,x_1} = \beta_0 + \beta_1 x_1 + \beta_3 x_3, \quad \mu_{Y|x_3,x_2} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each.

The one with the highest R^2 is chosen and compared to the reduced model $\mu_{Y|x_3} = \beta_0 + \beta_3 x_3$.

In this case it is the first model and we test

$$H_0: \beta_1 = 0.$$

We find that H_0 is rejected. Our model now includes x_1 .

Forward Selection Method

3. The three-variable model

$$\mu_{Y|x_1, x_2, x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is fitted and we test

$$H_0: \beta_2 = 0.$$

In this example, we find that H_0 can not be rejected. The final model is hence

$$\mu_{Y|x_3, x_1} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$$

Backward Elimination Procedure

Backward Elimination: One begins with a model that includes all the predictor variables and deletes them one at a time from the model until the reduced model is rejected.

28.3. Example. Assume that we have three potential predictor variables and that via backward elimination we obtain a reduced model containing only the variable X_2 . Assume that the variables X_1 and X_3 are deleted in this order. These are the steps taken:

1. The full model

$$\mu_{Y|x_1,x_2,x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is fitted. The value of R^2 is found.

Backward Elimination Procedure

2. The three two-variable models

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2,$$

$$\mu_{Y|x_1,x_3} = \beta_0 + \beta_1 x_1 + \beta_3 x_3,$$

$$\mu_{Y|x_2,x_3} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The model with the largest R^2 is chosen (here: $\mu_{Y|x_2,x_3}$) and compared with the full model. We test

$$H_0: \beta_1 = 0.$$

and are unable to reject H_0 . We hence delete X_1 from the model.

Backward Elimination Procedure

3. The two one-variable models

$$\mu_{Y|x_2} = \beta_0 + \beta_2 x_2, \quad \mu_{Y|x_3} = \beta_0 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The model with the largest R^2 is chosen (here: $\mu_{Y|x_2}$) and compared with the full model. We test

$$H_0: \beta_3 = 0.$$

and are unable to reject H_0 . We hence delete x_3 from the model.

4. We finally fit $\mu_Y = \beta_0$ and test

$$H_0: \beta_2 = 0.$$

and are able to reject H_0 . We hence keep x_2 and obtain the model
 $\mu_{Y|x_2} = \beta_0 + \beta_2 x_2$.

Stepwise Method

Stepwise Method: In forward selection, once a variable enters the model it stays. However, it is possible for one or more variables entering at a later stage to render a previously selected variable unimportant.

To detect this, each time a new variable enters in stepwise regression, all the variables in the previous model are checked for continued importance and possibly eliminated.

Hence, the stepwise method can be regarded as a combination of forwards election and backward elimination.

28.4. Example. In a multiple linear regression model, variables X_1 and X_3 are closely related, with variable X_1 being the best single predictor. Suppose that the final model contains the two variables X_2 and X_3 , with variable X_2 entering on the second stage. The steps in the stepwise regression are as follows:

Stepwise Method

1. The three single-variable models

$$\mu_{Y|x_1} = \beta_0 + \beta_1 x_1, \quad \mu_{Y|x_2} = \beta_0 + \beta_2 x_2, \quad \mu_{Y|x_3} = \beta_0 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The one with the highest R^2 is chosen and compared to the reduced model $\mu_Y = \beta_0$. In this case it is the first model and we test

$$H_0: \beta_1 = 0.$$

In this example, we find that H_0 is rejected. Our model now includes X_1 .

Stepwise Method

2. The two two-variable models

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2, \quad \mu_{Y|x_1,x_3} = \beta_0 + \beta_1 x_1 + \beta_3 x_3$$

are fitted. The value of R^2 is found for each. The one with the highest R^2 is chosen and compared to the reduced model

$\mu_{Y|x_1} = \beta_0 + \beta_3 x_1$. In this case it is the first model; we test

$$H_0: \beta_2 = 0.$$

and find that H_0 is rejected. We also check to see if X_1 is still needed, i.e., we test the model $\mu_{Y|x_1,x_2}$ for

$$H_0: \beta_1 = 0.$$

and reject H_0 . Thus X_2 alone is insufficient and our model now is

$$\mu_{Y|x_1,x_2} = \beta_0 + \beta_1 x_1 + \beta_2 x_2.$$

Stepwise Method

3. The three-variable model

$$\mu_{Y|x_1, x_2, x_3} = \beta_0 + \beta_1 x_1 + \beta_2 x_2 + \beta_3 x_3$$

is fitted; we test

$$H_0: \beta_3 = 0.$$

and reject H_0 . We now test whether X_2 is still needed,

$$H_0: \beta_2 = 0.$$

and reject H_0 . We also test whether X_1 is still needed,

$$H_0: \beta_1 = 0.$$

and fail to reject H_0 . Thus we eliminate X_1 and obtain the final model

$$\mu_{Y|x_2, x_3} = \beta_0 + \beta_2 x_2 + \beta_3 x_3$$

Never do this!

The above methods are commonly used, especially in **Data Mining**. However, the approach is actually **terrible**:

- ▶ We are performing **many** Fisher tests. Even disregarding all the problems with this type of test, the P -values are not accurate.

If we reject each hypothesis for $P < p_0$ and perform N **independent** tests, then the chance of having “falsely” (by our definition) rejected at least one of the H_0 a mistake is

$$(1 - p_0)^N$$

For large N , this can become quite large.

- ▶ But our tests are **not independent** in the first place - in fact, they are all performed on the same data set. That is a big problem, as we discussed earlier when talking about pre-tests.

Never do this!

- ▶ We are determining which tests to do based on data, rather than getting data based on pre-determined tests.
- ▶ The tests are biased to yield R^2 which is “too good” - the models are too well-fitted to the data, where the data itself may contain spurious features that disappear when new data is collected.
- ▶ The confidence intervals obtained from the data are too small. Also, often the final model is used as if it alone had been tested on the data, ignoring that previously lots of other models were discarded.

These and other issues are described in the web page cited below, where references to publications are also given.

Nowadays, there exist more sophisticated and improved approaches for model selection.

Literature:

<https://www.stata.com/support/faqs/statistics/stepwise-regression-problems/>

Regression to the Mean and Overfitting

In this context, it is a good idea to describe qualitatively ***Regression to the Mean***.

The basic idea is the following: if one performs two measurements of the same random variable and the first measurement result is very far away from the mean, then the second is likely to be closer to the mean.

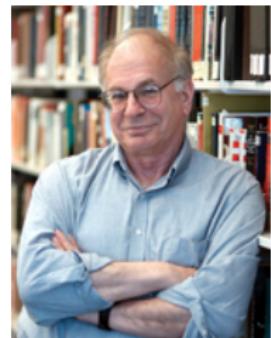
When fitting a model to a set of responses, certain responses may be far from their mean. Although the model may be a good fit for the given data, when using it for new data it may not fit as well, since the extreme responses have regressed to the mean.

This is an example of ***overfitting*** a model to a data set.

Kahnemann's Example

I had the most satisfying Eureka experience of my career while attempting to teach flight instructors that praise is more effective than punishment for promoting skill-learning.

When I had finished my enthusiastic speech, one of the most seasoned instructors in the audience raised his hand and made his own short speech [...].



Kahneman, D. in 2004 (1913–)
File:DanielKAHNEMAN.jpg. (2019, September 14). Wikimedia Commons, the free media repository.

He said, “On many occasions I have praised flight cadets for clean execution of some aerobatic maneuver, and in general when they try it again, they do worse. On the other hand, I have often screamed at cadets for bad execution, and in general they do better the next time. So please don’t tell us that reinforcement works and punishment does not, because the opposite is the case.”

Kahnemann's Example

This was a joyous moment, in which I understood an important truth about the world: because we tend to reward others when they do well and punish them when they do badly, and because there is regression to the mean, it is part of the human condition that we are statistically punished for rewarding others and rewarded for punishing them.

I immediately arranged a demonstration in which each participant tossed two coins at a target behind his back, without any feedback.

We measured the distances from the target and could see that those who had done best the first time had mostly deteriorated on their second try, and vice versa.

But I knew that this demonstration would not undo the effects of lifelong exposure to a perverse contingency.

David Kahnemann received the 2002 Nobel Memorial Prize in Economic Sciences for his work in behavioral economics and in the psychology of judgment and decision-making.

The Prediction Sum of Squares

There are various approaches to prevent overfitting. A basic idea is to test how well a model describes the existing data when the data points that it estimates are omitted.

A simple method is to calculate the ***prediction sum of squares (PRESS)*** for a model. This is done as follows:

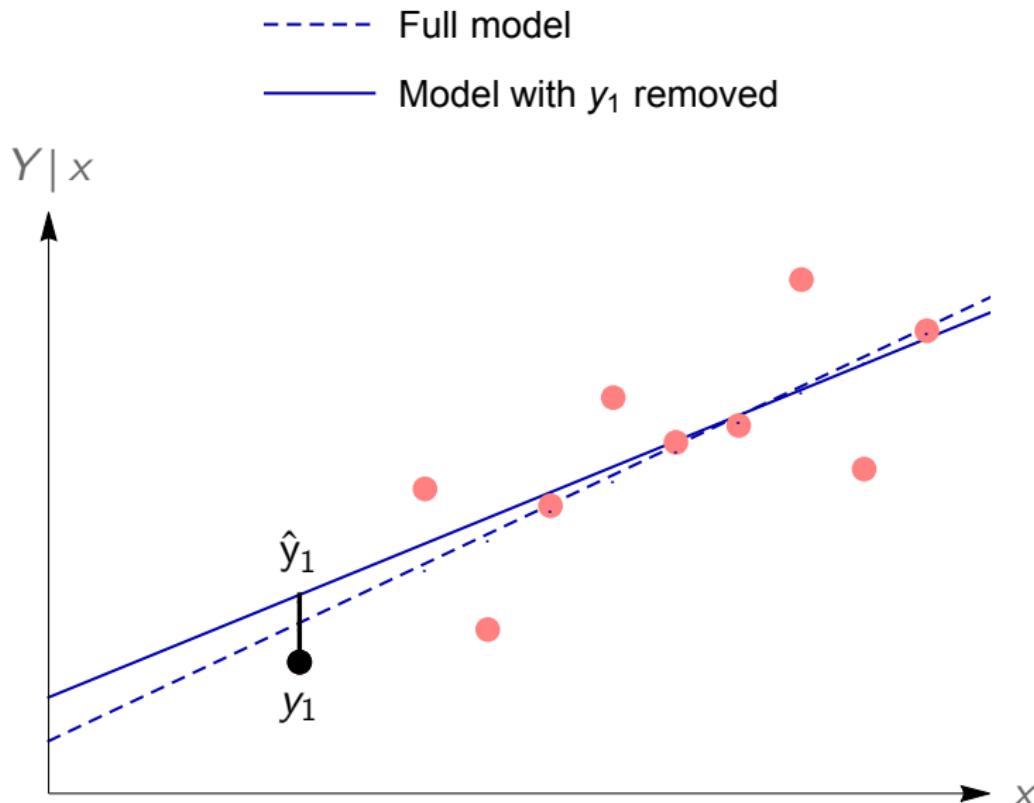
Given a model $Y | \mathbf{x}$ and a sample of size n , we calculate \hat{y}_i , $i = 1, \dots, n$, by omitting Y_i from the response data and fitting the model based on the remaining $n - 1$ data points.

We then calculate the ***PRESS statistic***

$$\text{PRESS} = \sum_{i=1}^n (y_i - \hat{y}_i)^2.$$

A small PRESS indicates that the model has not been fitted in a way that depends on extreme values of the responses.

The Prediction Sum of Squares



Concluding Remarks

Regression is today one of the most important tools of data science. Creating models and making inferences in fields such as machine learning, image recognition, behavior prediction and many other fields rely essentially on some type of regression.

However, finding the right model is hard. The last 50 years have seen many new and interesting approaches arise as old methods became subject to more intense scrutiny and were discarded.

No more than an introduction to the most basic concepts and methods is given here. We have not touched upon many issues of real practical interest, such as correlation between predictors and sophisticated techniques for model selection and comparison.

Nevertheless, hopefully this introduction has stimulated your interest in further investigations. Many specialized courses await!