1 Introduction

Statistics and Bioinformatics
Unversity of Zadar
Winter, 2011-2012

Contact Information

- Instructor: Stewart T. Schultz
  - email: sschultz@unizd.hr
  - phone: 200-653
- Teaching Assistants:
  - Melita Mokos
  - email: mmokos@unizd.hr
  - phone: 200-653
  - Ivana Zubak
  - email: izubak@unizd.hr
  - phone: 200-653
  - For administrative problems contact:
  - Tomislav Šarić
  - email: tosaric@unizd.hr
  - phone: 095 904 6339
- Course website: http://pomorstvo.unizd.hr/moodle/course/view.php?id=40
  - Username: zadar77
  - Password: zadar77

If you have any questions, feel free to contact any of us by email or phone.

Class schedule

- Lectures
  - Tuesday: 14:00 to 16:00 (two hours)
  - Thursday: 15:00 to 16:00 (one hour)
- Computer laboratory
  - Tuesday: 16:00 to 18:00 (two hours), SK-INF

Laboratory exercises will be using R, a programming language and an environment for statistical computing and graphics.

You will need to do your weekly quizzes outside the classroom at any computer with internet access (at home or on campus). You will need to have R installed on this computer.

To do this, go to http://www.r-project.org/, and click on "CRAN" on the left hand panel under downloads. Then choose a site to download from, perhaps from a site in Berlin, Germany. Then choose your operating system. Under Windows, choose the "base" distribution by clicking "install R for the first time". Then click "Download R 2.13.2 for Windows". This will give you a file with the extension .exe. Just click on it and it will install R on your computer. It will give you a blue "R" on your desktop, just click on that to start R running.

You will need to then install the package TeachingDemos within R. Open R, then go to the the menu item "Packages". Click "Install packages", then choose "Germany (Berlin)" and hit OK. Then scroll through the packages to "TeachingDemos", and select that, and hit OK. This will install TeachingDemos on your computer. Now go back to the R console and type "library(TeachingDemos)". Now TeachingDemos is active on your R session, and it will be active until you close R.

Course requirements

- Quizzes (15)
- Lab exercises (15)
- 3 midterm exams
- Optional final exam

Course grade

- Quizzes/Lab exercises: 50% of grade
- 3 midterm exams: 50% of grade

Course grade

- 60% required for a passing grade in **each** quiz, lab exercise, and midterm exam.
- Must attend at least 26 out of 30 lectures (allowed to miss up to four).
- Must attend at least 14 out of 15 labs (allowed to miss only one).
- If any of these requirements are not met, then you must take a cumulative final exam. This exam covers **all** the material in this course.
- If you do meet all requirements, then your grade will be based on 50% quiz/lab exercises, and 50% midterm exams.

Course grade

- You must take quizzes, lab exercises, and midterm exams online, on the Moodle website.
- A different quiz will be open on the website every week, and will be locked at Tuesday of the following week at midnight.
- You can take each quiz as many times as you want, in the one week window, before it is locked. Your grade on the quiz will be your highest grade of all attempts.

Course grade

Questions?

You learn statistics by doing statistics, therefore we have weekly quizzes, which you will complete online on the Moodle course page. You can take the quiz as many times as you want during the week it is open, and your final score on the quiz will be your maximum score. There will be a new quiz posted each week, and when the new quiz is posted the old quiz is locked and you can't take it anymore.

Also, since statistical analysis is done on the computer, you will do weekly exercises in the computer lab. These exercises will also be completed within Moodle, using R to answer the questions.

What is statistics?

- Set of tools
- to convert *data*
- to *probability*.

In this course, our data examples for the most part will come from agricultural research.

What is bioinformatics?

- Application of statistical tools
- to convert *molecular data*
- to *probability.*

Bioinformatics originally referred to information contained in any biological system, from genomes to ecosystems, but since 1980 it has referred exclusively to molecular data, primarily the analysis of sequences of DNA and amino acids.

2 Probability

What is probability?

- Axiomatic
  1. $0 \leq P(A) \leq 1$
  2. Impossible event: $P(A) = 0$
  3. Certain event: $P(A) = 1$
  4. Complement of A, A': $P(A') = 1 - P(A)$
  5. Two incompatible events A and B: $P(A \cup B) = P(A) + P(B)$
  6. Any two events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Relative frequencies
  - Repeat the experiment n times, tally the frequency of event A, n_A.
  - $$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

Both definitions of probability are useful, the axiomatic definition for simple symmetric systems and cases of random sampling, and the relative frequency definition within any experimental context. These axioms are easily visualized in a Venn Diagram or probability tree.

Conditional probability

Definition 1. The probability of A given $B \equiv P(A/B)$.

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

Thus,

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

Conditional probability simply recognizes the intuitive fact that if you know something about the context of an event, then its probability might change.

Independent events

Definition 2. Event A is independent of event B if and only if

$$P(A/B) = P(A)$$

Thus,

$$P(A \cap B) = P(A/B) \times P(B) = P(A) \times P(B)$$

"Product rule for independent events."

Event A is independent of event B if the probability of A is the same whether or not B occurs.

Bayes' Theorem

Theorem 3.

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

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

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

Bayes' Theorem allows calculation of the probability of event A conditioned on B, if you know the probability of event B conditioned on A. This is easily illustrated by a Venn Diagram or by a probability tree.

Screening test

Definition 4. A screening test is a preliminary test given to a sample of individuals to determine if they are likely to have a condition such as a disease, genotype, or contamination. Two events are analyzed: C, the presence of the condition; and T, a positive test result.

Screening test properties

Fundamental properties of a screening test T for a condition C:

- Positive predictive value: probability of having the condition given a positive test result; PPV = $P(C/T)$.
- Negative predictive value: probability of not having the condition given a negative test result; NPV = $P(C'/T')$.
- False positive rate: the probability of a positive test result given that the individual does not have the condition; $\alpha = P(T/C') = 1 -$ specificity.
- False negative rate: the probability of a negative test result given that the individual has the condition; $\beta = P(T'/C) = 1 -$ sensitivity.
- $$\text{Power} = 1 - \beta = \text{sensitivity}$$

Illustrate with a probability tree.

Fagan plot

Definition 5. An intuitive plot that relates the prior probability of a condition $P(C)$ to the posterior probability of a condition $P(C/T)$, based on the ratio of true positive rate to false positive rate, or power/α.

Lecture 2: Review of basic concepts of probability

3 Review of basic concepts of Probability

Lecture 3: Random variables and counting rules

Announcement

- Course schedule:
- Lectures:
- Monday 12:00 to 14:00 DHM

- Friday 13:00 to 14:00 DHM
-
- Laboratory exercises:
- Tuesday 14:00 to 15:00 SK-INF
- Tuesday 15:00 to 16:00 SK-INF

What is statistics?

Definition 6.
- Set of tools
- to convert *data*
- to *probability*.

Our tools so far

- Probability axioms
- Independence
- Conditional probability
- Bayes Theorem
- Venn Diagram, Probability Tree

What is probability?

- Axiomatic
  1. $0 \leq P(A) \leq 1$
  2. Impossible event: $P(A) = 0$
  3. Certain event: $P(A) = 1$
  4. Complement of A, A': $P(A') = 1 - P(A)$
  5. Two incompatible events A and B: $P(A \cup B) = P(A) + P(B)$
  6. Any two events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Relative frequencies
  - Repeat the experiment n times, tally the frequency of event A, n_A.
  - $$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

Axiomatic definition of probability

Definition 7.

$$P(A) = \frac{\text{number of outcomes giving event A}}{\text{total number of equally likely outcomes}}$$

Let's use some of these tools to do the Quiz ....

Another example

Bayes' Theorem, illustrated by screening test.

Screening nitrate with a test strip (Ross and Carlson, 1999). A white strip of paper turns pink or purple if there is more than a trace amount of nitrate in the sample. So a pink strip means high quality forage for goats or sheep. False positive rate is 0.98. False negative rate is 0.99. So it's a very good test, right? Well that depends on the prevalence, in this case the prevalence of plants with high nitrate content.

Emphasize: Definition of prevalance. Prevalence is just P(C), the probability of the condition in the population that you are interested. So in English we refer to the prevalence of AIDS, or the prevalence of breast cancer, or the prevalence of drug use, the prevalence of obesity, prevalence of alcoholism. All these are the overall probability that a random person drawn from the population has the condition.

The positive predictive value of the test is the proportion of all tests that are true positives.

The negative predictive value of the test is the proportion of all tests that are true negatives.

The sensitivity (or power) of the test is the proportion of all positive people who test positive. This is one minus the false negative rate. So sensitivity is P(T/C), where T is a positive test, and C is having the condition.

The specificity of the test is the proportion of all negative people who test negative. This is one minus the false positive rate. So specificity is P(T'/C').

The false positive rate of the test is the proportion of all negative people who test positive.

Other examples: The EMIT test, Enzyme Multiplied Immunoassay Test, for cocaine: $\alpha = 0.05$, power = 0.99, prevalence = 0.01.

False positive rate for opiates is 0.45 False positive rate for amphetamines is 0.8.

Example: mammography

power = 0.85 false positive rate = 0.05 prevalence in the population of breast cancer, for 40 yo women = 0.0064

Screening test in R

In R, we can construct a screening test table using:

- package TeachingDemos
- SensSpec.demo()

There are three ways to visualize conditional probability: probability trees, Venn diagrams, and 2×2 tables. The R demonstration constructs a full screening test table from three inputs: the sensitivity of the test, the specificity of the test, and the prevalence of the condition (or disease) in the population.

4 Random variables

Random variables

Definition 8. A random variable is a quantity whose value is not fixed, but which can take on different values according to a probability distribution.

Random variables

Example: number of boys in a family of three

Example: the number of boys in a family of three. There are four possibilities, 0, 1, 2, and 3, and each of these possibilities can be assigned a probability. In order to calculate these probabilities, first we need to tally up all the possible outcomes for families of three. Here are the possible outcomes: BBB, GGG, BGG, GBG, GGB, GBB, BGB, BBG. So there are eight possible outcomes. Now if we assume that each of these outcomes is equally likely, then we can calculate the probabilities.

Axiomatic definition of probability

Definition 9.

$$P(A) = \frac{\text{number of outcomes giving event A}}{\text{total number of equally likely outcomes}}$$

P(0) = 1/8 P(1) = 3/8 P(2) = 3/8 P(3) = 1/8

So: we have a random variable, let's call it x, the number of boys in a family of three children. There are four possible values of this variable. Each of these values has an associated probability.

Binomial expansion

This result can be illustrated with the binomial expansion.

$$(p+q)^3 = p^3 + 3p^2q + 3pq^2 + q^3.$$

Axiomatic definition of probability

Definition 10.

$$P(A) = \frac{\text{number of outcomes giving event A}}{\text{total number of equally likely outcomes}}$$

Fundamental counting principle

Definition 11. If a first task can be done in any one of m different ways and, after that task is completed, a second task can be done in any one of n different ways, then both tasks can be done, in the order given, in $m \times n$ different ways.

This can be illustrated with a probability tree diagram.

Permutations

Definition 12. A permutation is an ordered arrangement of a set of distinctly different items. For example, two permutations of the numbers one through five are 12345 and 53142.

Permutations

Theorem 13. *Through a direct application of the Fundamental Counting Principle, the number of permutations of n distinct items is given by*

$$_nP_n = n! = n(n-1)(n-2)\dots(2)(1)$$

The number of permutations of r items selected from n distinctly different items is

$$_nP_r = \frac{n!}{(n-r)!}$$

Example: 6 people seated in six chairs along a straight line. Example: 4 people out of 6 seated in four chairs along a straight line.

Combinations

Definition 14. A combination is an arrangement of a set of distinct items in which different orderings of the same items are considered the same. In other words, a combination is just a subset of a set of distinct elements.

Theorem 15. *The number of combinations of r items selected from n distinct items is*

$$_nC_r = \frac{n!}{r!(n-r)!} = \left(\begin{array}{c} n \\ r \end{array} \right)$$

Example: you go to the store to buy two shirts. The store has 3 different kinds of shirts. How many combinations of two shirts out of 3 can you buy? You can buy shirts 1 and 2, shirts 1 and 3, or shirts 2 and 3. So there are three possible combinations of two shirts selected out of a set of three shirts. Note that shirts 1 and 2 is the same combination as shirts 2 and 1, because order does not matter in this problem.

To understand the equation, we can arrange the possibilities under the general equation:

(number of combinations) × (number of permutations per combination) = total number of permutations

The total number of permutations is six: 12, 21, 13, 31, 23, 32. The number of permutations per combination is 2 (12 and 21, etc.). So the number of combinations is three.

We can then write down each of these: the number of permutations is 3×2. In terms of factorials, we can write this as 3!/1!. The number of permutations per combination is 2×1, which is 2!. So if we solve this equation for the number of combinations, we have

$$\text{number of combinations} = \frac{\text{total number of permutations}}{\text{number of permutations per combination}} = \frac{n!}{r!(n-r)!}$$

Laboratory exercises in R

- TeachingDemos: a package that illustrates statistical relationships
- sample() for random sampling from datasets
- SensSpec.demo() for constructing a screening test table

5 Review of counting rules and binomial distribution

Lecture 4: Review of counting rules and binomial distribution

What is statistics?

Definition 16.
- Set of tools
- to convert *data*
- to *probability.*

Our tools so far:

- Product rule for independence
- Bayes' theorem for the probability of a condition
- Binomial distribution for probability of "successes" in a binomial experiment

Binomial distribution is derived from

- Fundamental counting rule
- Number of permutations of n items taken r at a time
- Number of combinations of n items taken r at a time
- We assume independence!

Fundamental counting principle

Definition 17. If a first task can be done in any one of m different ways and, after that task is completed, a second task can be done in any one of n different ways, then both tasks can be done, in the order given, in $m \times n$ different ways.

Examples: Number of possible phone numbers beginning with a given prefix, say 200. Number of possible passwords, assuming either a letter or a digit, assuming say 26 letters and 10 digits, and letters can be either upper or lower case. Monkeys at a keyboard typing letters at random, what is the probability that a monkey types out Shakespeare's Hamlet? There are 199749 characters in Shakespeare's Hamlet.

Answer: around one in $37^{199749} = 10^{313246.7}$.

Permutations

Definition 18. A permutation is an ordered arrangement of a set of distinctly different items. For example, two permutations of the numbers one through five are 12345 and 53142.

Permutations

Theorem 19. *Through a direct application of the Fundamental Counting Principle, the number of permutations of n distinct items is given by*

$$_nP_n = n! = n(n-1)(n-2)\dots(2)(1)$$

The number of permutations of r items selected from n distinctly different items is

$$_nP_r = \frac{n!}{(n-r)!}$$

Examples: how many ways is it possible to seat 25 people in a straight line of chairs? Answer: 25!. How many distinct ways is it possible to arrange the 10 letters of the word *statistics*? Let's say we want to elect a president, vice-president, and secretary of our class. How many different ways can we do this?

Combinations

Definition 20. A combination is an arrangement of a set of distinct items in which different orderings of the same items are considered the same. In other words, a combination is just a subset of a set of distinct elements.

Theorem 21. *The number of combinations of r items selected from n distinct items is*

$$_nC_r = \frac{n!}{r!(n-r)!} = \left(\begin{array}{c} n \\ r \end{array} \right)$$

Examples: say we go shopping for shirts. At the store there are 10 shirts for sale and we want to buy 3. How many different combinations of 3 shirts out of 10 are possible? Let's say we want to buy two pairs of pants out of 15 that are available. How many combinations are possible? Now how many possible combinations of both pants and shirts are possible if we have 3 shirts and 2 pants, out of 10 and 15?

What is probability?

- Axiomatic
  1. $0 \leq P(A) \leq 1$
  2. Impossible event: $P(A) = 0$
  3. Certain event: $P(A) = 1$
  4. Complement of A, A': $P(A') = 1 - P(A)$
  5. Two incompatible events A and B: $P(A \cup B) = P(A) + P(B)$
  6. Any two events A and B: $P(A \cup B) = P(A) + P(B) - P(A \cap B)$
- Relative frequencies
  - Repeat the experiment n times, tally the frequency of event A, n_A.
  - $$P(A) = \lim_{n \to \infty} \frac{n_A}{n}$$

What is P(A)?

- If P(A) combines many incompatible outcomes a, b, c, d, etc., then
- $P(A) = P(a \cup b \cup c \cup d \cup ...) = P(a) + P(b) + P(c) + P(d) + ...$

If we want to know for example the probability of having 3 boys out of a family of six, then we need to find all the ways that this event can occur, find the probability of each of those ways, and add them up.

Probability of 3 boys out of 6 children?

1. Find all the way that this event can occur
2. Find the probability of each of these ways
3. Add them all up!

The number of ways that 3 boys out of a family of 6 can occur is the number of combinations of 3 out of 6. In R, you can type choose(6, 3), and this gives as an answer 20. The probability of each of these 20 ways is the probability of a boy taken to the third power, times the probability of a girl taken to the third power. We can multiply these probabilities together because we assume independence; i.e. we assume that the sex of one child does not affect the probability of the sex of subsequent children. So we are using the product rule for independence to get the probability of each possible way of getting 3 boys in a family of 6.

Probability of 3 boys out of 6 children?

$P(3 \text{ out of } 6) = {}_6C_3 p^{\text{boys}} q^{\text{girls}} = 20p^3q^3$, where p is the probability of a boy and $q = 1 - p$ is the probability of a girl.

We can generalize this to the Binomial distribution, for any number "n" of "trials" and any number "r" of successes among those trials. "Binomial" refers to the fact that there are just two possibilities, with no overlap: in our example we had boys or girls. "Trials" refers to the number of independent experiments in which we observed either one or the other outcomes. In our example each trial was a birth, and the outcome was boy or girl. "Successes" means that we often think of one outcome as a success (the one we're interested in) and the other outcome as a failure.

Binomial distribution

Theorem 22. *Let the random variable R denote the number of successes in n trials of a binomial experiment. Then R has a binomial distribution and the formula for the probability of observing exactly r successes in n trials is given by*

$$P(R = r) = \left(\begin{array}{c} n \\ r \end{array} \right) p^r q^{n-r} \text{for } r = 0, 1, 2, \ldots n, \text{ where}$$

n = number of trials, r = number of successes in n trials, p = probability of success in a single trial, $q = 1 - p$, and $\left(\begin{array}{c} n \\ r \end{array} \right) = {}_n C_r = \frac{n!}{r!(n-r)!}$

In R, we can calculate a binomial probability very easily. Say we want to know the probability of 10 successes in 20 trials, with the probability of success in one trial equal to 0.2. Then the probability of 10 successes in 20 trials is

```
choose(20, 10)*0.2^10*0.8^10
```

The answer we get is 0.107.

Note that we assume that each trial is independent of each other trial. This allows us to use the product rule of independence, and multiply all the probabilities together to compute the final probability.

6 Random variables: Binomial and Poisson

Lecture 5: Random variables

Announcements:

- Lecture Mondays 12:00 to 14:00, Fridays 13:00 to 14:00
- Lab exercises Tuesdays 14:00 to 16:00
- Quizzes 1 and 2 end tomorrow (Tuesday October 25)!

Lecture 5: Random variables

Today's schedule:

- Random variables
- Probability distributions
- Measures of central tendency, spread
- Binomial distribution
- Poisson distribution

What is statistics?

Definition 23.

- Set of tools
- to convert *data*
- to *probability*.

What is data?

Definition 24. Data is a collection of *random variables* that may be related in some way.

Examples of random variables

- Number of female children in families of 5
- Number of diseased plants in a population of 100
- Number of deaths in a population of 100 over some time period
- Photosynthesis rate of plants
- Respiration rate of plants
- Harvest dry weight of plants
- Number of seeds germinating out of 100
- etc.

Random variables

Definition 25. A random variable is a quantity whose value is not fixed, but which can take on different values according to a **probability distribution**. A probability distribution consists of a chart, table, graph, or formula that specifies all the values that the random variable can assume along with the probability of those values occurring.

Examples: we take a random sample of students and measure height or weight. A random sample of plants and measure seet set, fruit set, harvest yield, growth rate, survivorship from seed to harvest. Or the number of offspring in a family with a trait or condition of interest, such as disease, or disease resistance, or growth rate, or some trait we're interested in.

Example: flipping a coin once: the probability distribution is 0.5 on heads and 0.5 on tails. Rolling a die once: the probability distribution is 1/6 on each of the six possibilities. THe complication here comes when you flip a coin more than once. Now how do we distribute probabilities over 1, 2, 3, 4, etc. heads?

Draw a graph of a probability distribution to visualize the distribution of probability across possible values of the variable.

There are two types of probability distribution the probability mass distribution for discrete variables, and the probability density distribution for continuous variables

Probability distributions

Two kinds:

- Probability mass function for a discrete variable
- Probability density function for a continuous variable

Probability distributions

Two fundamental properties:

- Location
- Spread

Draw pictures of probability distributions varying in location and spread.

Measures of location, or "central tendency"

- Mode
- Median
- Mean, or "Expected value" of x, $E(x) = \mu$

Mean of a distribution

Definition 26. The mean of a discrete probability distribution is

$$E(x) = \mu = \sum_i x_i P(x_i).$$

The mean of a continuous probability distribution is

$$E(x) = \mu = \int_{-\infty}^{+\infty} x f(x) dx.$$

In both cases, the mean is the "balance point" of the distribution.

Measures of variability

- Variance $= Var(x) = \sigma_x^2$
- Standard deviation $= \sqrt{Var(x)} = \sigma_x$

Variance of a distribution

Definition 27. The variance of a discrete probability distribution is

$$Var(x) = E(x-\mu)^2 = \sum_i (x_i - \mu)^2 P(x_i).$$

The variance of a continuous probability distribution is

$$Var(x) = E(x-\mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx.$$

Properties of the mean

- $E(x+c) = E(x) + c$
- $E(ax) = aE(x)$
- $E(x+y) = E(x) + E(y)$, even if x and y are dependent (correlated).
- $E(xy) \neq E(x)E(y)$
- $E(x^2) \neq [E(x)]^2$

Properties of the variance

- $Var(ax) = a^2 Var(x)$
- $Var(x+y) = Var(x) + Var(y)$ only if x and y are independent (not correlated)

Let's apply to the binomial distribution

Axiomatic definition of probability

Definition 28.

$$P(A) = \frac{\text{number of outcomes giving event A}}{\text{total number of equally likely outcomes}}$$

$$P(A) = (\text{number of possible outcomes giving event A})$$
$$\times(\text{probability of one outcome giving event A})$$

Binomial distribution

Definition 29. Consider n independent trials of an experiment where at each trial there are exactly two possible outcomes: success with probability p that is constant from trial to trial and failure with probability $1-p$. Let X equal the number of successes out of the n repeated trials. The random variable X is called a **binomial random variable**, and its probability distribution is called a **binomial distribution**.

Mean and variance of the binomial distribution

Theorem 30. *If x is a binomial random variable with n trials and probability of success p, then the mean of this variable is np and its variance is $np(1-p) = npq$.*

Binomial distribution

Theorem 31. *Let the random variable X denote the number of successes in n trials of a binomial experiment. Then X has a binomial distribution and the formula for the probability of observing exactly x successes in n trials is given by*

$$P(X = x) = \begin{pmatrix} n \\ x \end{pmatrix} p^x q^{n-x} \text{for } x = 0, 1, 2, \ldots n, \text{ where}$$

x = number of trials, x = number of successes in n trials, p = probability of success in a single trial, $q = 1 - p$, and $\begin{pmatrix} n \\ x \end{pmatrix} =_n C_x = \frac{n!}{x!(n-x)!}$

Poisson distribution

The Poisson distribution can be thought of a special case of the binomial distribution for very large n and very small p.

In that case, the mean np and variance $np(1-p)$ are essentially equal to each other.

Poisson distribution

The probability of exactly k events during some interval or across some spatial region, given that the mean number of events is λ, is equal to

$$P(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

Binomial distribution in R

- dbinom(0:10, 10, 0.5) gives the probability of each of zero to 10 successes in a sample of 10, where the probability pf one success is 0.5
- pbinom(0:10, 10, 0.5) gives the cumulative probability of each of zero to zero to 10 successes in a sample of 10, where the probability pf one success is 0.5.
- plot(0:10, dbinom(0:10, 10, 0.5)) plots the above on a graph, which gives a visual depiction of the binomial probability distribution for those values.

Poisson distribution in R

- dpois(0:10, 2) gives the probability of each of zero to 10 events across some spatial or temporal sampling area, where mean number of events in that area is 2
- ppois(0:10, 2) gives the cumulative probability of each of zero to 10 events across some spatial or temporal sampling area, where mean number of events in that area is 2.
- plot(0:10, dpois(0:10, 2)) plots the above on a graph, which gives a visual depiction of the poisson probability distribution for those values.

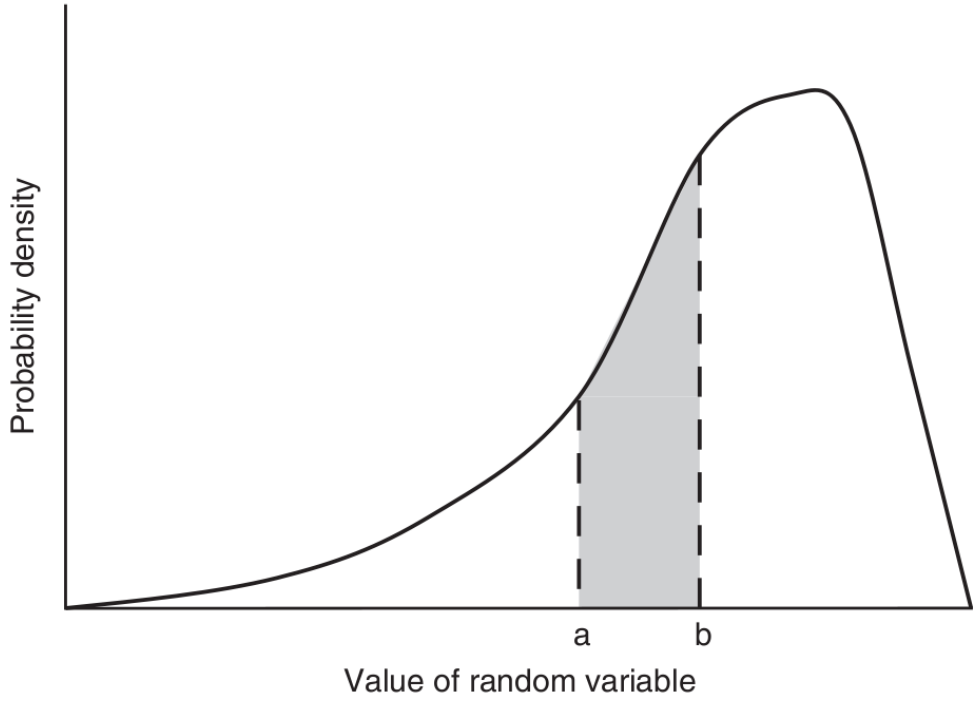
7 Random variables: Review

Lecture 6: Random variables: review

What is statistics?

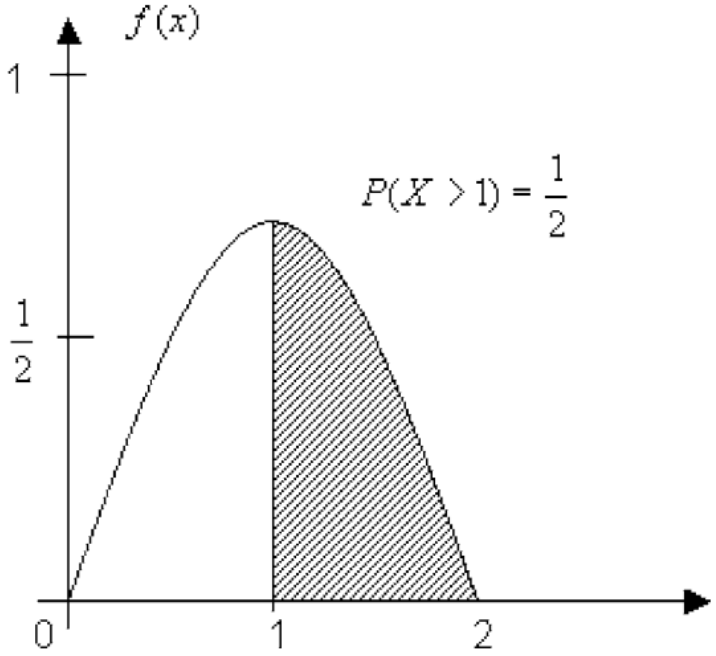
Definition 32.
- Set of tools
- to convert *data*
- to *probability*.

Probability distributions



Probability distributions

In this probability distribution, the probability that x is greater than one equals one-half:



Probability distributions

Two fundamental properties:

- Location

- Spread

Measures of location, or "central tendency"

- Mode
- Median
- Mean, or "Expected value" of x, $E(x) = \mu$

Mean of a distribution

Definition 33. The mean of a discrete probability distribution is

$$E(x) = \mu = \sum_i x_i P(x_i).$$

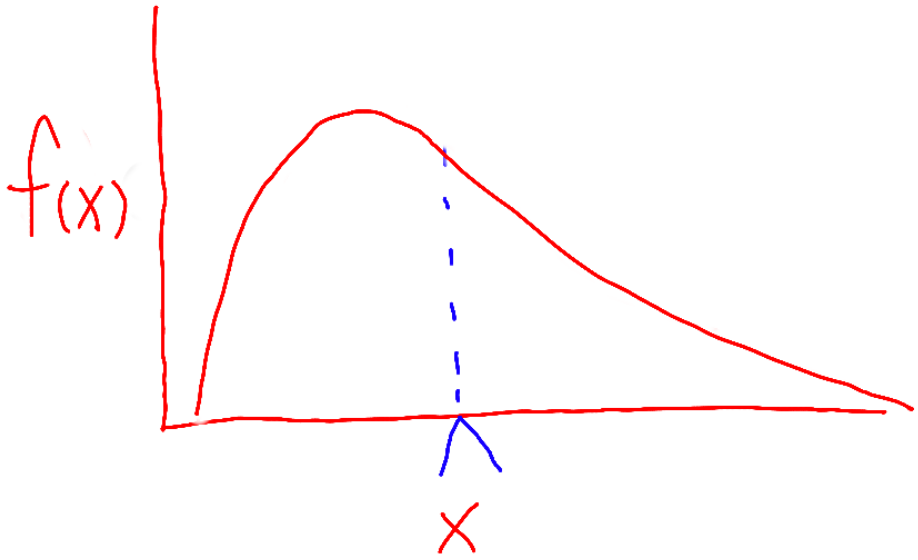
The mean of a continuous probability distribution is

$$E(x) = \mu = \int_{-\infty}^{+\infty} x f(x) dx.$$

In both cases, the mean is the "balance point" of the distribution.

Mean of a distribution

The mean is the balancing point of the distribution, if we treat the area under the curve as physical mass.



Measures of variability

- Variance $= Var(x) = \sigma_x^2$
- Standard deviation $= \sqrt{Var(x)} = \sigma_x$

Variance of a distribution

Definition 34. The variance of a discrete probability distribution is

$$Var(x) = E(x-\mu)^2 = \sum_i (x_i - \mu)^2 P(x_i).$$

The variance of a continuous probability distribution is

$$Var(x) = E(x-\mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx.$$

Properties of the mean

- $E(c) = c$, where c is a constant.
- $E(x + c) = E(x) + c$, where c is a constant.
- $E(cx) = cE(x)$, where c is a constant.
- $E(x + y) = E(x) + E(y)$, even if x and y are dependent (correlated).
- $E[E(x), E(y)] = E(x, y)$,
- $E(xy) \neq E(x)E(y)$
- $E(x^2) \neq [E(x)]^2$

Binomial distribution

- We have n trials, and the probability of success on each trial is p.
- p is a constant.
- Each trial is independent of every other trial (one trial does not affect another trial).
- Mean of the binomial distribution $E(x) = np$.
- Variance of the binomial distribution $Var(x) = np(1-p) = npq$.

Binomial distribution

Probability of r successes in n trials:

$$\begin{aligned} P(r, n) &= \text{number of ways r successes} \times \text{probability of one way} \\ &= \text{number of ways to get r successes} \times p^r q^{n-r} \\ &= \text{combinations of n items taken r at a time} \times p^r q^{n-r} \\ &= \frac{n!}{r!(n-r)!} \times p^r q^{n-r} \end{aligned}$$

Statistical test

We want to know if some observation can be explained by chance variation, or if some additional cause is present.

1. We suppose that we have a known probability distribution.
2. We suppose that our observation was drawn from that probability distribution.
3. We ask: what is the probability that we would have obtained a value this extreme (or more) from this distribution?
4. If this probability is less than 0.05, then we conclude that this observation could not (likely) have occurred by chance; i.e. it is "unusual" or "unexpected." Therefore, there must have been some cause outside the variation in the given probability distribution.
5. If this probability is greater than or equal to 0.05, then we conclude that this observation could have occurred by chance.

Binomial distribution in R

- dbinom(0:10, 10, 0.5) gives the probability of each of zero to 10 successes in a sample of 10, where the probability pf one success is 0.5
- pbinom(0:10, 10, 0.5) gives the cumulative probability of each of zero to zero to 10 successes in a sample of 10, where the probability pf one success is 0.5.
- plot(0:10, dbinom(0:10, 10, 0.5)) plots the above on a graph, which gives a visual depiction of the binomial probability distribution for those values.

Examples: number of lymphocytes in a sample of blood cells. Number of a particular type of offspring. Number of individuals dying over a given time interval (also might be a Poisson distribution), number of disease infected individuals in a population.

Poisson distribution

- Poisson distribution is the binomial distribution with large n and small p.
- If $n > 100$ and $p < 0.01$, and all the assumptions of the binomial are met, then we have the Poisson distribution.
- Then the mean is equal to the variance.
- $\sigma^2 = Var(x) = np(1-p) \cong np = mean(x) = \mu$.
- Poisson distribution is appropriate for events in space or time, where one event is independent of other events.
- Examples: genetic mutation, dispersion of individuals in space, morphology, weather events (e.g. storms), accidents, number of deaths in a region or across a time interval, number of parasites found on a host, radioactive decay.

Poisson distribution in R

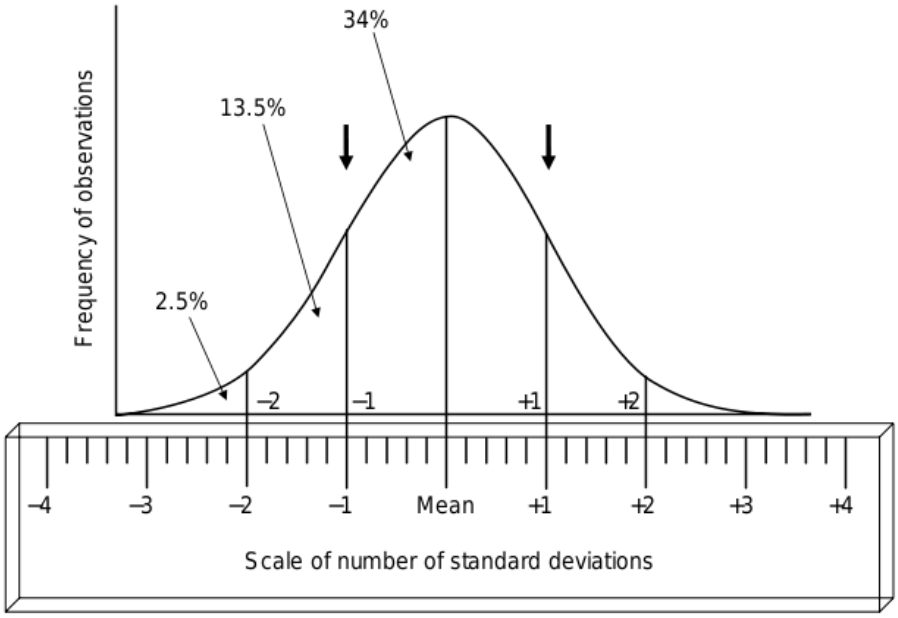
- dpois(0:10, 2) gives the probability of each of zero to 10 events across some spatial or temporal sampling area, where mean number of events in that area is 2
- ppois(0:10, 2) gives the cumulative probability of each of zero to 10 events across some spatial or temporal sampling area, where mean number of events in that area is 2.
- plot(0:10, dpois(0:10, 2)) plots the above on a graph, which gives a visual depiction of the poisson probability distribution for those values.

Examples from the binomial distribution: can a person predict the outcome of a coin flip? Do this in class. Another example: lymphocytes in a sample of 100 blood cells. Probability of HIV positive is 0.008. What if 500 babies are screened, and 5 are found positive? Number of people getting cancer over a particular time interval. Let's say the probability that someone dies between age 20 and age 40 is 0.002. Let's say we have 100 students age 20 in a class, and they have their 20-year reunion, and 3 people died. Is that an unusual class?

Examples from the Poisson distribution: let's say there is an expected average of 1 new mutation per seed. We screen a seed for mutations, and we find 3 new mutations. Is this expected purely by chance? Number of buras per winter is 10. This winter there were 20. Is this expected by chance, or could some other factor be involved, maybe climate change? Let's say the average number of offspring of a particular bird is 3. We take a particular bird we're interested in and find that it produces 6 offspring. Could this have happened by chance, or is some other factor involved (perhaps an unusual diet)?

Normal distribution

Fundamental rule: approximately 95% of all observations lie within 2 standard deviations of the mean!



8 Normal distribution

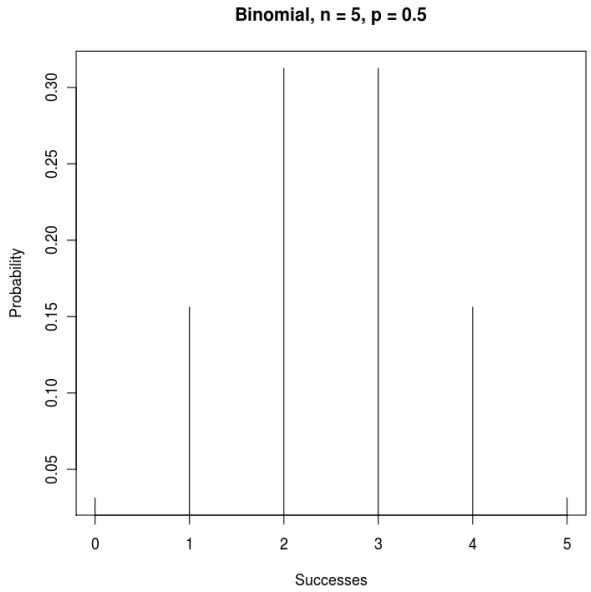
Lecture 7: Normal distribution and the Central Limit Theorem

Reading: pp. 11 - 44 in Vasilj.

What is statistics?

Definition 35.
- Set of tools
- to convert *data*
- to *probability.*

Probability distributions

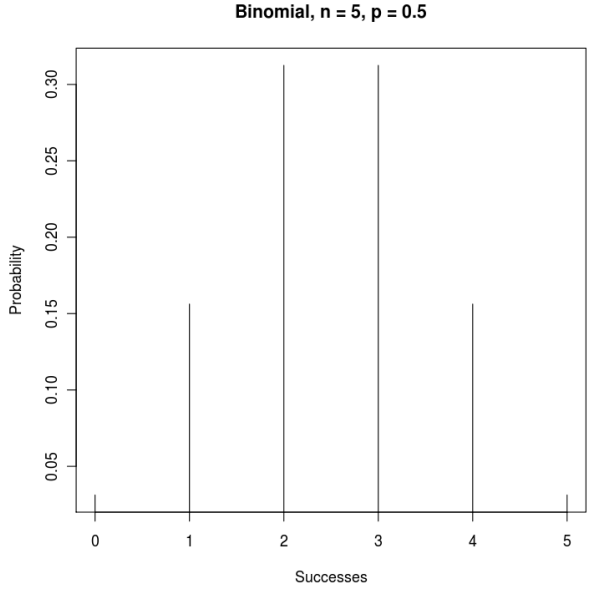


plot(0:5, dbinom(0:5, 5, .5), type = "h", xlab = "Successes", ylab = "Probability", main = "Binomial, n = 5, p = 0.5")

Binomial distribution

- We have n trials, and the probability of success on each trial is p.
- p is a constant.
- Each trial is independent of every other trial (one trial does not affect another trial).
- Mean of the binomial distribution $E(x) = np$.
- Variance of the binomial distribution $Var(x) = np(1-p) = npq$.

Binomial distribution



$\mu = E(x) = np = 5 \times 0.5 = 2.5$.
$\sigma^2 = Var(x) = E(x-\mu)^2 = np(1-p) = 5 \times 0.5 \times 0.5 = 1.25$.
$\sigma = \sqrt{\sigma^2} = \sqrt{1.25} = 1.12$.

Binomial distribution: Example

- Given: probability that a blood cell is a lymphocyte is 0.22 in a healthy human population.
- In one patient's blood test, 100 cells were counted and 32 were lymphocytes.
- Is this normal?
- Need to calculate $P(X = 32)$ or more from a binomial distribution with $n = 100$ and $p = 0.22$.
- $P(X = r) = \frac{n!}{r!(n-r)!} \times p^r q^{n-r}$
- $P(0 \leq X \leq r) =$ pbinom(r, n, p)
- $P(r \leq X \leq n) = 1 -$ pbinom(r-1, n, p)

We do this with the pbinom() function. pbinom(r, n, p) gives the probability that the number of successes lies between 0 and r inclusive. So to get the probability that the number of successes is between r and n inclusive, we compute 1 - pbinom(r-1, n, p). This follows from the axioms of probability.

Statistical test

We want to know if some observation can be explained by chance variation, or if some additional cause is present.

1. We suppose that we have a known probability distribution.
2. We suppose that our observation was drawn from that probability distribution.
3. We ask: what is the probability that we would have obtained a value this extreme (or more) from this distribution?
4. If this probability is less than 0.05, then we conclude that this observation could not (likely) have occurred by chance; i.e. it is "unusual" or "unexpected" or "significant." Therefore, there must have been some cause outside the variation in the given probability distribution.
5. If this probability is greater than or equal to 0.05, then we conclude that this observation could have occurred by chance.

Binomial distribution: Example 2

- Given: probability that a coin lands heads is 0.5.
- I flip a coin 10 times, and a student guesses correctly on 4 flips.
- Can the student predict a coin flip?
- Need to calculate $P(X = 4)$ or better from a binomial distribution with $n = 10$ and $p = 0.5$.
- 1 - pbinom(3, 10, 0.5)

Poisson distribution

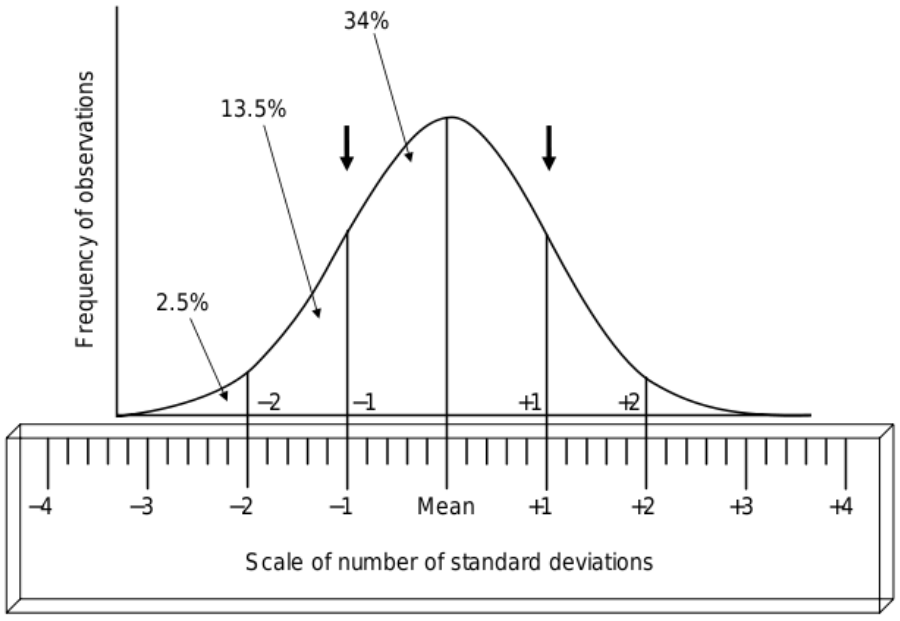
- Poisson distribution is the binomial distribution with large n and small p.
- If $n > 100$ and $p < 0.01$, and all the assumptions of the binomial are met, then we have the Poisson distribution.
- Then the mean is equal to the variance.
- $\sigma^2 = Var(x) = np(1-p) \cong np = mean(x) = \mu$.
- Poisson distribution is appropriate for events in space or time, where one event is independent of other events.
- Examples: genetic mutation, dispersion of individuals in space, morphology, weather events (e.g. storms), accidents, number of deaths in a region or across a time interval, number of parasites found on a host, radioactive decay.

Poisson distribution: Example 1

- Given: in a normal plant, the mean number of mutations per genome each generation is 1.0.
- I expose a plant to a chemical, then measure 3.0 mutations in one seed.
- Is the chemical a mutagen?
- Need to calculate $P(X = 3)$ or more from a Poisson distribution with $\lambda = 1.0$.
- $P(X = k) = \frac{\lambda^k e^{-\lambda}}{k!}$.
- 1 - ppois(2, 1)

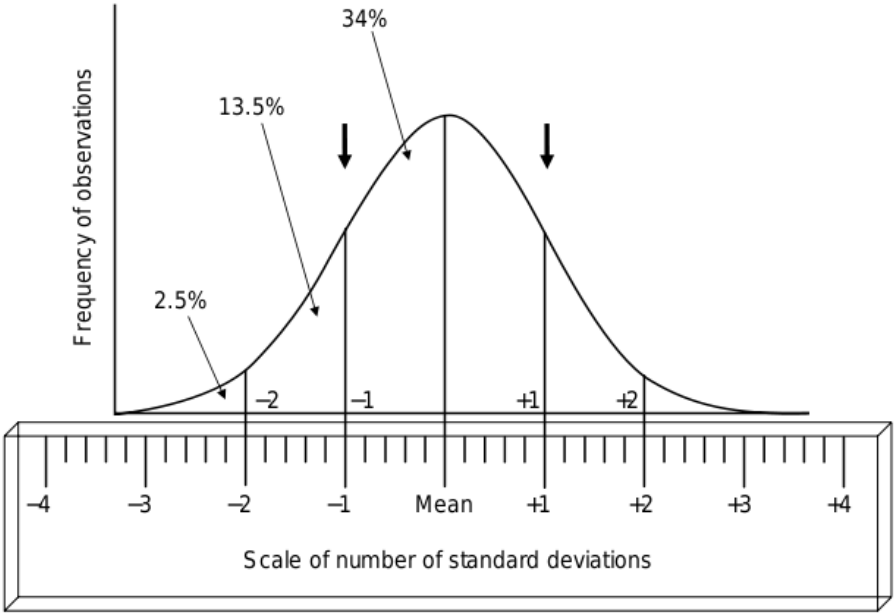
Normal distribution

- The normal distribution is bell-shaped.
- The normal distribution is symmetrical.
- Approximately 95% of all observations lie within 2 standard deviations of the mean.



Normal distribution: Example 1

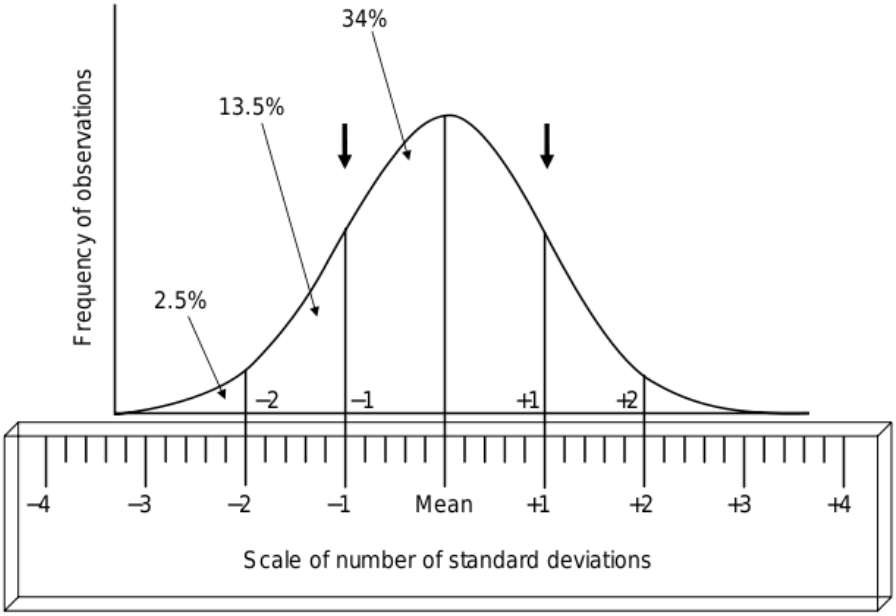
- The mean height of an adult Arabidopsis plant is 15 cm, and the standard deviation of height is 2 cm.
- I randomly sample a seed from a mutant variety of Arabidopsis. I grow it to adult, and its height is 10 cm.
- Does the mutant gene cause stunted growth?
- Need to calculate $P(X = 10)$ or lower, assuming a normal distribution.
- pnorm(10, 15, 2)



Note that the normal distribution is continuous, while the Binomial and Poisson distributions are discrete. For this reason, you don't have to worry about the probability of the endpoint of the interval you are calculating. So, for the binomial probability of more than 6 successes out of 10 trials, for p = 0.5, you compute 1 - pbinom(6-1, 10, 0.5). We have to subtract 1 from six, otherwise we would count the probability of 6 twice. For the normal distribution probability of greater than 6 if the mean is 5 and standard deviation 1.58, we compute 1 - pnorm(6, 5, 1.58). We don't have to subtract 1 from six, because the probability of six is zero, because this is a continuous distribution.

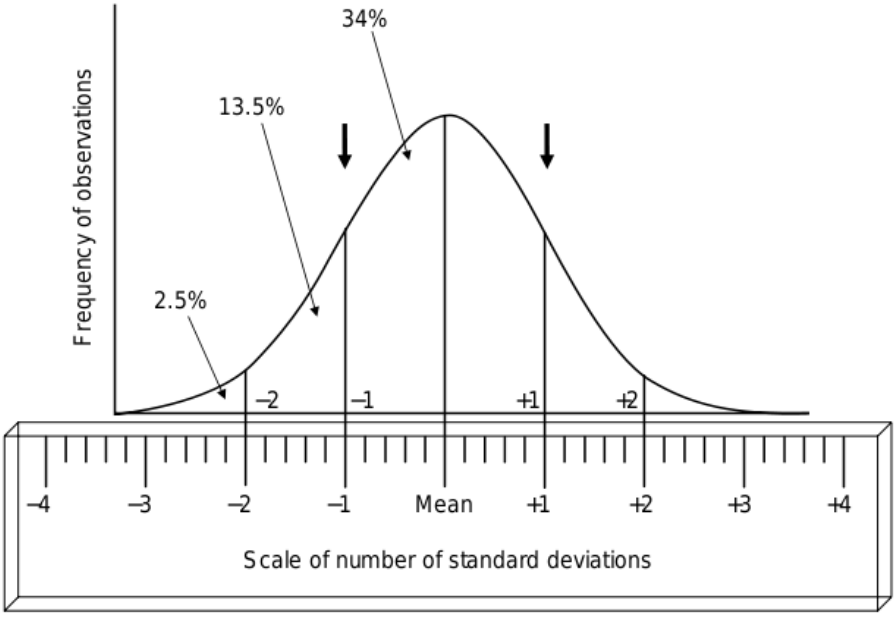
Normal distribution

- The normal distribution is bell-shaped.
- The normal distribution is symmetrical.
- Approximately 95% of all observations lie within 1.96 standard deviations of the mean.



Normal distribution

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



Where did this equation come from?

Central Limit Theorem

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

Theorem 36. *The sum of n identically-distributed variables call S_n. Then S_n approaches a normal distribution as n approaches infinity.*

Central Limit Theorem

The central limit theorem can be demonstrated in a Galton board, or a Quincunx machine:
`http://www.youtube.com/watch?v=xDIyAOBa_yU`
`http://www.mathsisfun.com/data/quincunx.html`

Normal distribution

The binomial distribution becomes the normal distribution if $np(1-p) > 5$. Compare the two when $n = 16$ and $p = 0.5$:

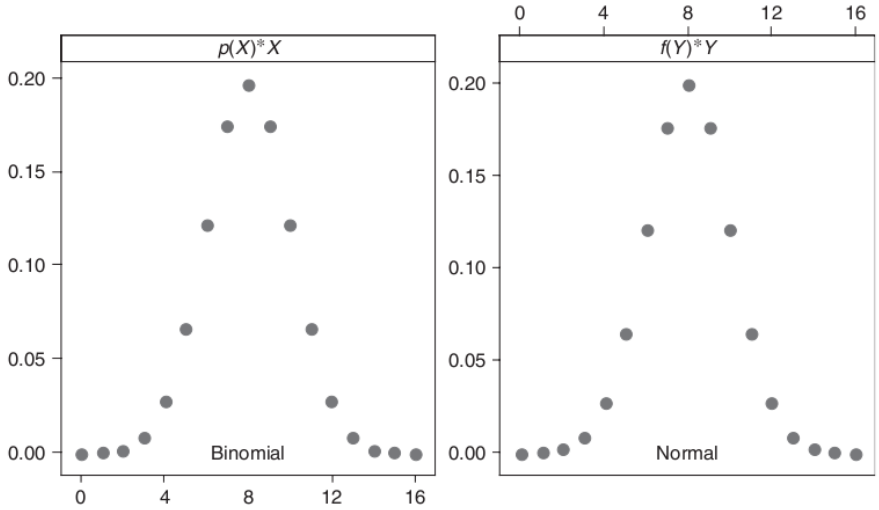
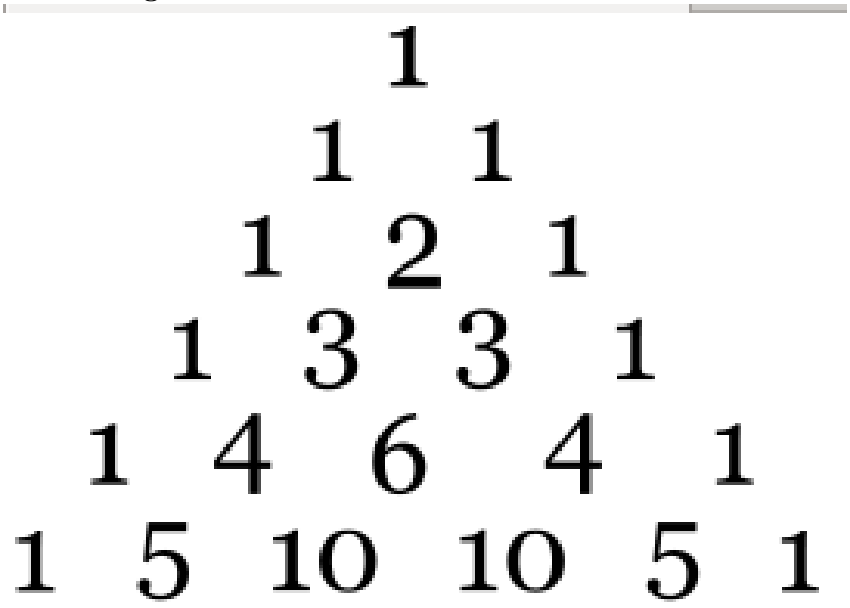


Fig. 7-2 Plot of binomial for $N = 16$ and $p = 0.5$ and normal curve with mean = 8 and standard deviation = 2.

Here, $np(1-p) = (16)(0.25) = 4$, so we wouldn't usually consider this quite a normal distribution, but it is very very close. If $n = 20$, then it is virtually indistinguishable from normal.

Similarly, the Poisson distribution becomes a normal distribution when $\lambda > 5$. Here, the variance is also greater than 5, because the variance and mean are equal.

Pascal's triangle



The normal distribution can be illustrated with Pascal's triangle. Each row represents a biological experiment with n trials, with n = the number of rows counted down from the top. Each number in the triangle represents the number of ways of getting r successes, where r is equal to the number of slots in the row, counted in from the left. Pascal's triangle itself is illustrated by the Quincunx machine.

9 Central Limit Theorem continued

Lecture 8: Central limit theorem, normal distribution continued

Announcements:

- Reading for today's subject: pp. 48 - 52 in Vasilj.

Central Limit Theorem

The central limit theorem can be demonstrated in a Galton board:

`http://www.youtube.com/watch?v=xDIyAOBa_yU`

`http://www.mathsisfun.com/data/quincunx.html`

Any repeated random process eventually generates the normal distribution. In this case the repeated random process is a ball bouncing against pegs, left to right with probability 0.5. In other words a binomial variable with n equal to the number of rows on the board, and p equal to any constant.

Central Limit Theorem

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

Theorem 37. *The sum of n identically-distributed variables call S_n. Then S_n approaches a normal distribution as n approaches infinity.*

Question: What if the variables are not identically-distributed? See the Galton board to see what happens.

Answer: we can calculate the sample mean!

The sample mean is calculated $\bar{x} = \frac{\sum_i^n x_i}{n}$.

Question: what is the shape of the distribution of the sample mean?

Answer: We know that the sample mean should approach a normal distribution, by the Central Limit Theorem.

If you are drawing from two or more distributions, then the resulting sum is not necessarily normally distributed. You must be drawing from a single distribution, at random, every time, for the sum to be normally distributed.

Central Limit Theorem

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

Theorem 38. *The sum of n identically-distributed variables call S_n. Then S_n approaches a normal distribution as n approaches infinity.*

Corollary: the sample mean of these n identically-distributed variables is normally distributed.

Question: what is the mean of the sample mean?

Corollary: the mean of the sample mean is the mean of the distribution.

Question: what is the variance of the sample mean?

Corollary: the variance of the sample mean of the n variables is their variance divided by n. The standard deviation of the sample mean of the n variables is the standard deviation divided by $\sqrt{n}$.

Again, the central limit theorem requires that all the variables that we sum have the same "identical" distribution, identical mean and variance. If they don't, then the sum is no longer normally distributed.

This can be demonstrated on the Quincunx machine on the mathisfun website, by changing the p-value. Let it run with p=0.5 for awhile, then switch to p=.1 or something, and see what happens. You get a bimodal distribution.

Are we then out of luck? Can we convert this to a normal distribution somehow? Yes we can, by taking a random sample from it. Every time we take a sample from it, we have a sample from the same "identical" distribution. The distribution does not change between samples. So if we take a random sample from that distribution, then get the sum of that sample, the sum is normally distributed, by the central limit theorem.

But, if the sum is normally distributed, then the sample mean also is. The sample mean is just $\bar{x} = \frac{\sum_i x_i}{n}$, where n is the sample size. Notice that the numerator is the sum of many identically distributed variables (we know they are all identically distributed because we have sampled randomly from the same distribution), so the numerator is likely to be normally distributed if n is large enough. Dividing by n is just dividing by a constant, so that has no effect on the distribution, so the sum divided by n is also normally distributed. So the sample mean is normally distributed.

This is a very useful result. By the central limit theorem, we know that the sample mean will be approximately normally distributed, and it will be closer to the normal distribution as n, the sample size, increases. Generally, if we are sampling from a distribution that is already close to a normal distribution, such as height or mass of individuals, then we need exceed only a small n in order for the mean to be normally distributed. Usually $n > 20$ or so is enough sample size to assume the mean is normal.

Draw a normal distribution, show its mean, now imagine that we take random samples of $n = 100$ or some other number from this population, and for each sample we calculate the mean and mark it on the graph. Will our estimates of the mean be as spread out as the original distribution? No, their distribution will be narrower. In other words, the distribution of the sample mean has a lower variance than the distribution of teh population you are sampling from. How much lower? We can calculate this from the properties of the variance that we demonstrated last week.

Properties of the mean

- $E(ax) = aE(x)$
- $E(x + y) = E(x) + E(y)$ for all variables x and y with finite variance

Central Limit Theorem

The expected value of the sample mean is the mean of the distribution, $E(\bar{x}) = E(\sum_i^n x_i/n) = E(x) = \mu$.

Proof. We know that $E(x+y) = E(x)+E(y)$. Also, $E(ax) = aE(x)$. Then $E(x_1+x_2) = E(x_1)+E(x_2) = 2E(x)$. Then $E(x_1 + x_2 + x_3 + \ldots + x_n) = nE(x)$. Or, equivalently, $E(\sum_i^n x_i) = nE(x)$. But we want to know $E(\sum_i^n x_i/n)$. This is just the sum divided by the constant. Therefore the mean is the sum of the means divided by the constant, $E(\bar{x}) = E(\sum_i^n x_i/n) = nE(x)/n = E(x) = \mu$. □

Properties of the variance

- $Var(ax) = a^2Var(x)$
- $Var(x + y) = Var(x) + Var(y)$ only if x and y are independent (not correlated)

Central Limit Theorem

Theorem 39. [...]

Corollary: the variance of the sample mean of the n variables is their variance divided by n. The standard deviation of the sample mean of the n variables is the standard deviation divided by $\sqrt{n}$.

Central Limit Theorem

Proof. We know that $Var(x + y) = Var(x) + Var(y)$, if x and y are independent. Also, $Var(ax) = a^2Var(x)$. Then $Var(x_1 + x_2) = Var(x_1) + Var(x_2) = 2Var(x)$. Then $Var(x_1 + x_2 + x_3 + \ldots + x_n) = nVar(x)$. Or, equivalently, $Var(\sum_i^n x_i) = nVar(x)$. But we want to know $Var(\sum_i^n x_i/n)$. This is just the sum divided by the constant. Therefore the variance is the sum of the variances divided by the constant squared, $Var(\bar{x}) = Var(\sum_i^n x_i/n) = nVar(x)/n^2 = Var(x)/n$. Then the standard deviation is the square root of the variance: $SD(\bar{x}) = \sqrt{Var(\bar{x})} = \sqrt{Var(x)/n} = SD(x)/\sqrt{n} \equiv$ S.E.M. □

Central Limit Theorem

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

Theorem 40. *The sum of n identically-distributed variables call S_n. Then S_n approaches a normal distribution as n approaches infinity.*

Corollary: the sample mean of these n identically-distributed variables is normally distributed.

Corollary: the variance of the sample mean of the n variables is their variance divided by n. The standard deviation of the sample mean of the n variables is the standard deviation divided by $\sqrt{n}$.

Central Limit Theorem

In summary:

- The mean of all the sample means is μ, the population mean.
- The standard deviation of all the sample means is $\frac{\sigma}{\sqrt{n}}$, where σ is the population standard deviation.
- The variance of all the sample means is $\frac{\sigma^2}{n}$, where σ^2 is the population variance.

Central Limit Theorem

General procedure if we know the true standard deviation σ:

- Given we know the true mean μ and true standard deviation σ from the distribution.
- We have an observation x_i.
- We calculate the Z score of our observation:

$$Z = \frac{x_i - \mu}{\sigma}.$$

- This is the distance from the mean to the observation in units of standard deviation.
- If this distance is in the tail of the distribution (e.g. if $|Z| > 1.96$) then we conclude that the observation is "unusual" or "unexpected" or "significant".
- Or, we can calculate the probability that the observation is at least as extreme as observed.

Example. Let's imagine that a hotel owner has a hotel with 9 rooms. He has a boiler that can give 90 minutes of hot water for the 9 showers. He wants to know what is the probability that if he has all 9 guests showering at the same time, they will use up all the 90 minutes of water? Well that would require that the average shower length for those 9 guests is 10 minutes. So, what's the probability that the average (sample mean) shower length for 9 random guests is 10 minutes? To calculate this we need to know the mean and standard deviation. Let's say he knows (he has been keeping track over several months) that the mean shower length of people in his hotel is 8 minutes, and the standard deviation is 3 minutes. What is the probability that the mean shower length of 9 people would be higher than 10 minutes? The standard deviation of the mean, or S.E.M. is 3 divided by the square root of 9, or $3/3 = 1$. The mean is 8 minutes. So 10 minutes is 2 standard deviations higher than 8, and 2 standard deviations is 95% of the normal distribution. So our Z-score is 2.0, and the probability that 9 people all showering at the same time would use up all the water is approximately 0.025, or 2.5%.

NOTE our assumptions here. We assume the sample mean is a normal distribution. Can we assume this? The reason we assume it is that we have a sample of 9. Is 9 large enough of a sample to assume normality? It might be, if the variable of shower time is close to normal, as you can see from the R TeachingDemo clt.examp(). Just adding together 5 or so uniform random variates is enough to give a distribution close to normal. But in practice, we might be a little concerned that the distribution is normal if we have a sample size only of 9. We also assume that we already know the mean and standard deviation of shower time of the hotel guests, and that the 9 people are a random sample from all potential guests of this hotel. If they are not a random sample, then we can't make any conclusions. For example, the mean and standard deviation may have changed recently because people changed their showering habits, etc., so now we have a sample from a different distribution than that assumed.

10 Estimation

Lecture 9: Estimation with confidence intervals

Announcements:

- Reading for today's subject: pp. 52 - 54 in Vasilj.
- Midterm 1 is next week.
- Midterm 1 Practice exam is now online.
- The practice exam is very similar to the Midterm Exam 1.
- Today's subject: Z scores, estimation, and confidence intervals.

Lecture 9: Estimation with confidence intervals

Remember, the standard deviation of the sample mean is

$$\frac{\sigma}{\sqrt{n}}$$

- This is also called the SEM or standard error of the mean.
- We can use this to create error bars around our estimate of the mean.
- There are two conventions for error bars: the simpler one is that the error bar is just the standard error of the mean.
- The more complicated one is that the error bar is the "confidence interval" around the mean.
- Both can be better understood in terms of Z, the standard normal variate.

Z scale

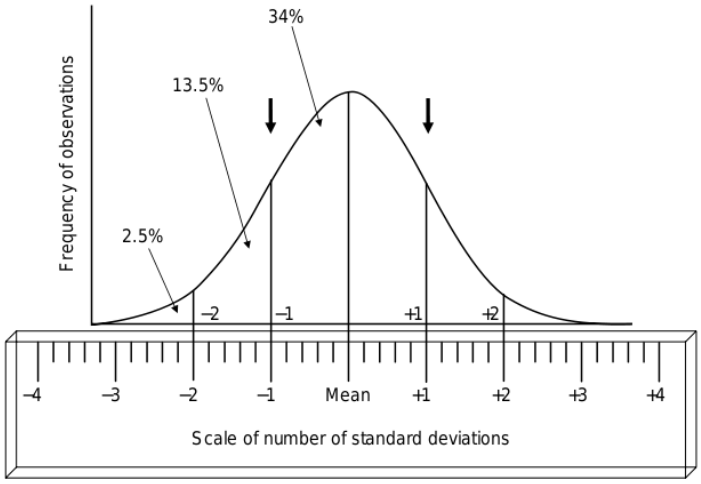
General procedure if we know the true standard deviation:

- Given we know the true mean μ and true standard deviation σ from the distribution.
- We have an observation x_i.
- We calculate the Z score of our observation:

$$Z_i = \frac{x_i - \mu}{\sigma}.$$

- This is the distance from the mean to the observation in units of standard deviation.
- If this distance is in the tail of the distribution (e.g. if $|Z_i| > 1.96$) then we conclude that the observation is unusual.
- $Z_{0.025} = 1.960$.
- $Z_{0.05} = 1.645$.
- $Z_{0.005} = 2.576$.

Z distribution



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

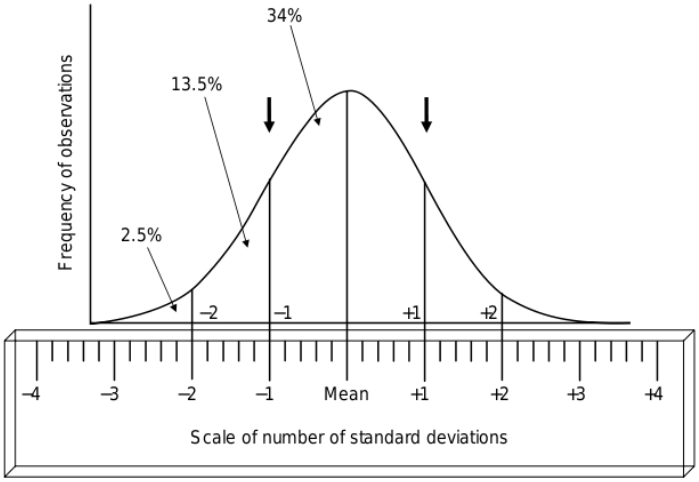
$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Notice that the mean of the Z variable is 0, and the standard deviation and variance are both equal to 1.

Z distribution in R: pnorm()

For any Z_i score, the probability that Z is less than or equal to Z_i,

$$P(Z \leq Z_i) = \text{pnorm}(Z_i).$$

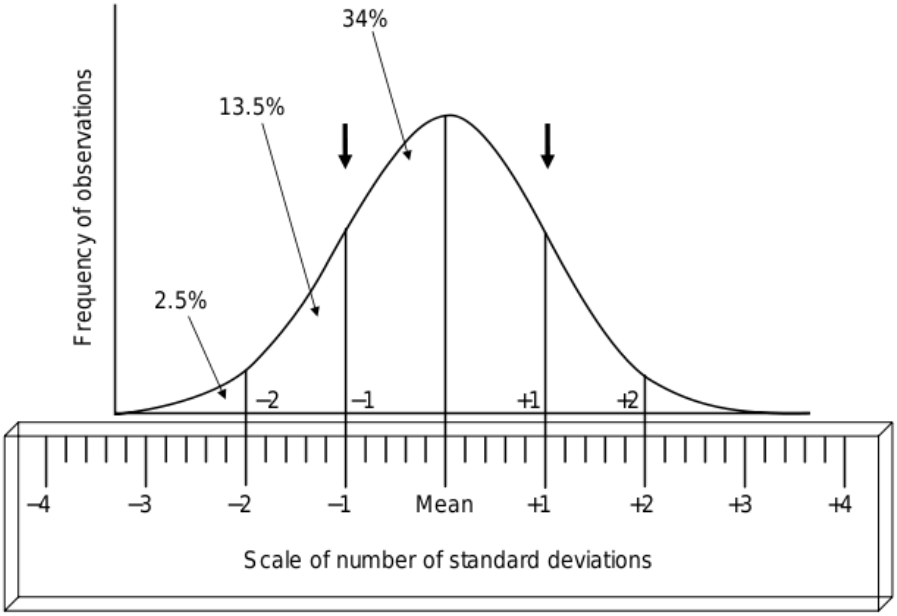


You should know (memorize) that the Z variable is normally distributed with mean 0 and standard deviation 1. The Z distribution is called the "standard normal" distribution, i.e. the distribution with mean 0 and standard deviation 1. All normal distributions can be converted to the Z distribution; in fact it makes them easier to work with. It's just changing the scale.

Example: assume the mean height of men is 175 cm, and the standard deviation is 8 cm, and height is normally distributed. Then my height of 183 cm has a Z score of 1, it is one standard deviation higher than the mean. In R, the probability that a random man is shorter than I am is pnorm(1). The probability that a random man is taller is 1-pnorm(1). pnorm() assumes that the mean is zero and the

standard deviation is 1 unless you tell it otherwise. So, for example, pnorm(1) is the same as pnrom(183, mean=175, sd = 8) since 183 has a Z score of 1 if the mean is 175 and the sd is 8. So in general, for any Z score, pnorm(Z) is the probability of Z or smaller.

Z distribution



Z critical values

$$P\left(-2.576 \leq Z \leq 2.576\right) = 0.99.$$
$$P\left(-1.960 \leq Z \leq 1.960\right) = 0.95.$$
$$P\left(-1.645 \leq Z \leq 1.645\right) = 0.90.$$

Notice that most of the time the normal distribution that we will be working with is for the sample mean. So we can convert a sample mean into a Z value. But for the sample mean, the standard deviation is not σ, rather the standard deviation is $\sigma/\sqrt{n}$. Remember the standard deviation of the mean is the SEM, the standard error of the mean. So the Z value for the sample mean is $(\bar{x} - \mu)/(\sigma/\sqrt{n})$.

Confidence interval

Definition 41. A confidence interval is any interval that we can say with confidence contains a parameter of interest. The parameter of interest is usually the population mean μ or variance σ^2. Our level of confidence is determined by the method of constructing the interval. The method will produce an interval containing the estimated parameter with a probability equal to the level of confidence.

Calculating the 95% confidence interval

$$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95.$$
$$P\left(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95.$$
$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x} - \mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

$$P\left(-\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Calculating the 95% confidence interval

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

This is the 95% confidence interval for the sample mean.
We are 95% confident that the true mean is contained in this interval.

$$Z_{0.025} = 1.96.$$

$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

For example, we want to estimate the mean height of Croatian men. We take a random sample of 5 men, measure their height. Let's say the mean is 169. What is the 95% confidence interval? It is (162, 176). Let's say we take another sample of 20 men, and find that the sample mean is 176. Then the 95% confidence interval is (172.5, 179.5). Let's say we take another sample of 100 mean, and find the sample mean is 174. Then the 95% confidence interval is (172.4, 175.6). Let's say we take a random sample of 1000 men, and get a mean of 175.3. Now the 95% confidence interval is (174.8, 175.8). You should do these calculations yourself to check your knowledge. As you can see, as the sample size increases, the width of the confidence interval gets smaller. In other words, our estimate becomes closer and closer to the true mean, and our 95% confidence of the true mean extends over a smaller and smaller segment of possible values.

We don't have to calculate the 95% confidence interval, we could also calculate the 90% confidence interval, or the 99% confidence interval, or some other interval. It depends on how certain we would like to be about the true value of the population mean. The 99% confidence interval is wider than the 95% and 90% confidence intervals.

This procedure can be followed for any sample mean from any distribution, as long as n, the sample size is large enough that we can assume the normal distribution. The calculation can be done for the binomial distribution and for the Poisson distribution. The mean of a binomial variable (the number of successes out of n trials, with probability of success p), is just the observed proportion of successes, $n_{\text{successes}}/n$.

Lecture 10: Estimation with confidence intervals, continued

Announcements:

- Reading for this week's subject: pp. 54 - 79 in Vasilj.
- No lab exercises next Tuesday.
- Instead Midterm Exam 1 will be during the lab period.
- You need to sign up for the exam. Sign up for one of three sessions:

  14:00 to 14:50

  15:00 to 15:50

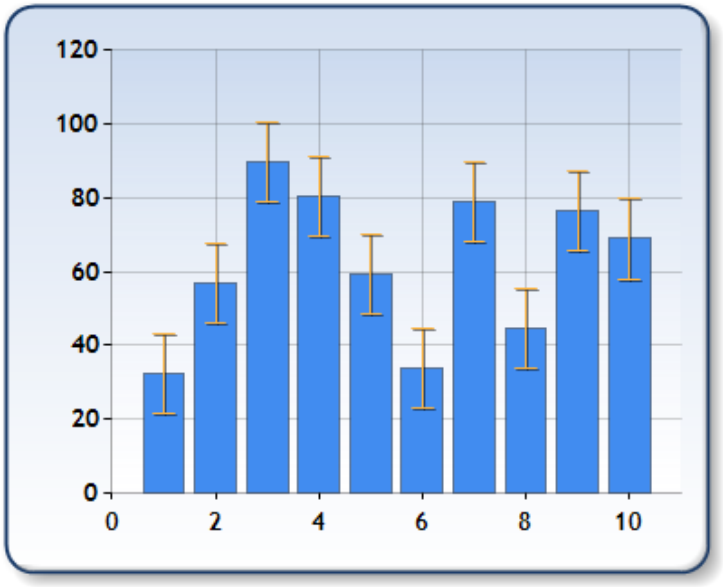
  16:00 to 16:50

Lecture 10: Estimation with confidence intervals, continued

Rules for Midterm Exam 1

- You have 50 minutes.
- This exam is closed book, closed notes, open mind.
- You are allowed one side of one sheet of paper for any notes you want.
- You may take the exam as many times as you want, within the 50 minute time limit.
- Questions?

Lecture 10: Estimation with confidence intervals



Population versus sample

Definition 42. A **population** is the set of all persons or things of interest. A **sample** is a subset of the population. Convention: Greek letters are used for population parameters (μ for mean, σ for standard deviation. Latin letters are used for the sample ($\bar{x}$ for sample mean, s for sample standard deviation). We say $\bar{x}$ is an **estimate** of μ, and s is an estimate of σ.

Show a diagram of population and sample, with corresponding probability distributions. Show the definition of mean of the probability distribution, and the definition of mean of the sample.

We can illustrate the effects of sample size on the distribution of the mean with R, using the dnorm() function. Say for heights of men, the distribution will have a mean of 175 cm and standard deviation of 8 cm. To make a plot, we type xv = seq(140, 215, .1) to get the x values to plot, and then plot the distribution with plot(xv, dnorm(xv, mean=175, sd = 8), type="l"). We can show the effect of sample size with points(xv, dnorm(xv, mean=175, sd = 8/sqrt(n)), where n is the sample size.

For example: IQ scores for adults are normally distributed, with mean $\mu = 100$ and standard deviation $\sigma = 15$. If a random sample of 16 adults is selected, find the probability that the average IQ score in the sample will be at least 92.5. Here the standard deviation of the mean is $15/4 = 3.75$, and 92.5 is $(-7.5)/3.75 = -2$ on the Z scale. So the probability that the mean is at least -2 on the Z scale is 1 - pnorm(-2) = 0.977.

We can use the standard error of the mean (SEM) to construct the margin of error around the estimate of the true mean.

Draw another picture of the normal distribution, and show that 95% of the distribution is between plus or minus 1.96 standard deviations.

As I showed last time, the 95% confidence interval is easily calculated:

Calculating the 95% confidence interval

$$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95.$$

$$P\left(-1.96 \leq \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95.$$

$$\text{Therefore, } P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

This is the 95% confidence interval for the sample mean.

$$Z_{0.025} = 1.96.$$

$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

Calculating the 95% confidence interval of the sample mean $\bar{x}$

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (\bar{x} - E, \bar{x} + E)$$

where

1 - α = confidence coefficient (1 - α)100% = level of confidence α = the error probability $\bar{x}$ = the mean of the sample σ = the standard deviation of the population n = the sample size $Z_{\alpha/2}$ = the critical value of Z, or the Z value that cuts off a right-tail area of $\alpha/2$ under the standard normal curve; that is, $1 - P(Z < Z_{\alpha/2}) = \alpha/2$ $Z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$ = E, the margin of error.

Calculating the 95% confidence interval of the sample mean $\bar{x}$

$$\bar{x} \pm Z_{\alpha/2}\frac{\sigma}{\sqrt{n}} = (\bar{x} - E, \bar{x} + E)$$

Table 10.1 Commonly Used Critical Values $z_{\alpha/2}$

Level of Confidence $(1-\alpha)100\%$	Critical Value $z_{\alpha/2}$
90%	1.645
95%	1.96
99%	2.576

As an example, let's say there is a new diet for people who want to lose weight. We sample 100 people who use the diet for two months, and we keep track of their weight gain during that time. The sample mean weight gain is -1 kg, and the standard deviation weight gain is 7 kg. Does the diet work? The standard error of the mean is 7/10 kg. Thus the mean, -1, is less than 2 standard errors from zero, so it is not unexpected assuming the true mean is zero. So we cannot conclude that the diet works.

The sample standard deviation

Definition 43. The sample variance is written s^2 and is defined as:

$$s^2 = \frac{\sum_1^n (x_i - \bar{x})^2}{n-1}.$$

The sample standard deviation is just the square root of the sample variance:

$$s = \sqrt{s^2} = \sqrt{\frac{\sum_1^n (x_i - \bar{x})^2}{n-1}}.$$

$n-1$ is called the "degrees of freedom" in this calculation.

General rule (to be revisited later)
If the sample size is 30 or greater, we assume Z can be calculated using the sample standard deviation s rather than the population standard deviation σ.

11 Z tests and t tests

Lecture 11: Estimation with Z and t distributions
Announcements:

- Reading for this week's subject: pp. 54 - 79 in Vasilj.
- Midterm Exam 1 tomorrow 14:00 to 18:00
- Please signup for a time slot if you haven't already.

Lecture 11: Estimation with Z and t distributions
Rules for Midterm Exam 1

- You have 50 minutes.
- This exam is closed book, closed notes, closed internet, closed Wikipedia, closed past quizzes, open mind.
- You are allowed one side of one sheet of paper for any notes you want.
- You may take the exam as many times as you want, within the 50 minute time limit.
- Questions?

Lecture 11: Estimation with Z and t distributions
On the menu for today

- Estimation continued with the Z distribution
- ... using the normal approximation from the Binomial
- ... using the normal approximation from the Poisson
- Estimation continued with the t distribution

Z scale

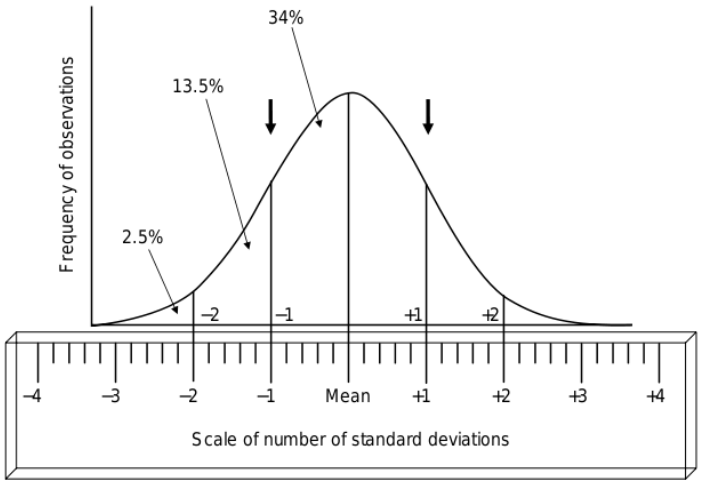
General procedure if we know the true standard deviation:

- Given we know the true mean μ and true standard deviation σ from the distribution.
- We have an observation x_i.
- We calculate the Z score of our observation:

$$Z_i = \frac{x_i - \mu}{\sigma}.$$

- This is the distance from the mean to the observation in units of standard deviation.
- If this distance is in the tail of the distribution (e.g. if $|Z_i| > 1.96$) then we conclude that the observation is unusual.
- $Z_{0.025} = 1.960$.
- $Z_{0.05} = 1.645$.
- $Z_{0.005} = 2.576$.

Z distribution



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

$$f(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Z distribution

To convert the sample mean to the Z-scale:

$$Z_{\bar{x}} = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

To convert the sample sum to the Z-scale:

$$Z_{\sum x} = \frac{\sum x - n\mu}{\sqrt{n\sigma^2} = \sigma\sqrt{n}} = \frac{\sum x - n\mu}{\sigma\sqrt{n}}$$

Population versus sample

Definition 44. A **population** is the set of all persons or things of interest. A **sample** is a subset of the population. Convention: Greek letters are used for population parameters (μ for mean, σ for standard deviation. Latin letters are used for the sample ($\bar{x}$ for sample mean, s for sample standard deviation). We say $\bar{x}$ is an **estimate** of μ, and s is an estimate of σ.

Calculating the 95% confidence interval

$$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95.$$
$$P\left(-1.96 \leq \frac{\bar{x}-\mu}{\frac{\sigma}{\sqrt{n}}} \leq 1.96\right) = 0.95.$$
$$P\left(-1.96\frac{\sigma}{\sqrt{n}} \leq \bar{x}-\mu \leq 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$
$$P\left(-\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \leq -\mu \leq -\bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$
$$P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

Calculating the 95% confidence interval

$$P\left(\bar{x}-1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x}+1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$
$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

This is the 95% confidence interval for the sample mean.

We are 95% confident that the true mean is contained in this interval.

$$Z_{0.025} = 1.96.$$
$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

Normal distribution

The binomial distribution becomes the normal distribution if $np(1-p) > 5$. Compare the two when $n = 16$ and $p = 0.5$:

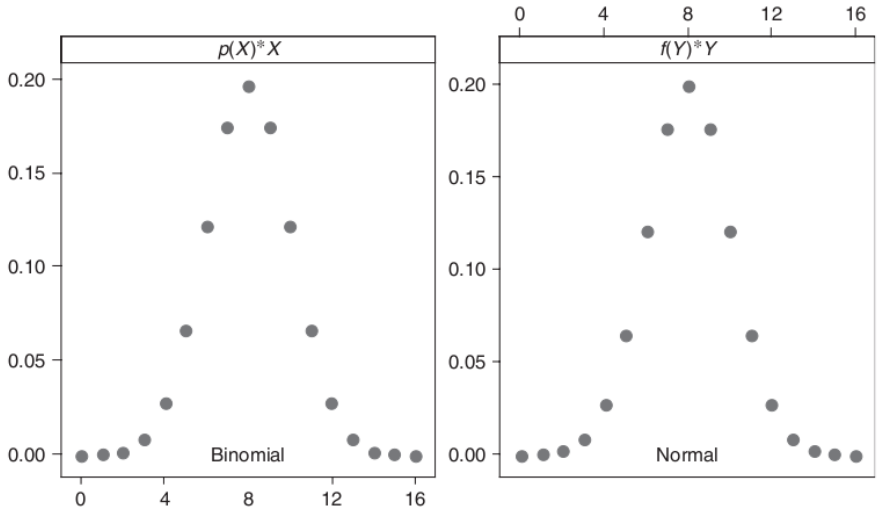


Fig. 7-2 Plot of binomial for $N = 16$ and $p = 0.5$ and normal curve with mean = 8 and standard deviation = 2.

Mean and variance of the binomial distribution

Theorem 45. *If x is a binomial random variable with n trials and probability of success p, then the mean of this variable is $\mu = np$ and its variance is $\sigma^2 = np(1-p) = npq$. Its standard deviation is $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$.*

Z scale for the binomial distribution

Theorem 46. *If x is a binomial random variable with n trials and probability of success p, then the mean of this variable is $\mu = np$ and its variance is $\sigma^2 = np(1-p) = npq$. Its standard deviation is $\sigma = \sqrt{npq} = \sqrt{np(1-p)}$.*

Let's say we observe s successes in n trials of a binomial experiment. What is s on the Z scale?

$$Z_s = \frac{s-\mu}{\sigma} = \frac{s-np}{\sqrt{np(1-p)}}.$$

Calculating the 95% confidence interval from the binomial distribution

$$P\left(-1.96 \leq Z \leq 1.96\right) = 0.95.$$

$$P\left(-1.96 \leq \frac{s-np}{\sqrt{np(1-p)}} \leq 1.96\right) = 0.95.$$

$$P\left(-1.96\sqrt{np(1-p)} \leq s-np \leq 1.96\sqrt{np(1-p)}\right) = 0.95.$$

$$P\left(-s-1.96\sqrt{np(1-p)} \leq -np \leq -s+1.96\sqrt{np(1-p)}\right) = 0.95.$$

$$P\left(s-1.96\sqrt{np(1-p)} \leq np \leq s+1.96\sqrt{np(1-p)}\right) = 0.95.$$

Calculating the 95% confidence interval from the binomial distribution

$$P\left(s-1.96\sqrt{np(1-p)} \leq np \leq s+1.96\sqrt{np(1-p)}\right) = 0.95.$$

$$s \pm 1.96\sqrt{np(1-p)}$$

This is the 95% confidence interval for the sample mean.
We are 95% confident that np is contained in this interval.

$$Z_{0.025} = 1.96.$$

$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

Mean and variance of the proportion in a binomial experiment

Theorem 47. *If x is a binomial random variable with n trials and probability of success p, then we can calculate the proportion of successes as $\hat{p} = x/n$. What is the mean of this proportion? $E(\hat{p}) = \mu = p$. What is the variance of this proportion?*

$$Var(\hat{p}) = \sigma^2 = Var\left(\frac{x}{n}\right) = \left(\frac{1}{n^2}\right)Var(x) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

What is the standard deviation of this proportion?

$$\sigma = \sqrt{\sigma^2} = \sqrt{\frac{p(1-p)}{n}}$$

Calculating the 95% confidence interval from the binomial distribution

$$P\left(\hat{p} - 1.96\sqrt{p(1-p)/n} \leq p \leq \hat{p} + 1.96\sqrt{p(1-p)/n}\right) = 0.95.$$

$$\hat{p} \pm 1.96\sqrt{p(1-p)/n}$$

This is the 95% confidence interval for the true probability of success p.

We are 95% confident that p is contained in this interval.

$$Z_{0.025} = 1.96.$$

$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

Poisson distribution

The Poisson distribution can be thought of a special case of the binomial distribution for very large n and very small p.

In that case, the mean np and variance $np(1-p)$ are essentially equal to each other.

Poisson distribution

The probability of exactly k events during some interval or across some spatial region, given that the mean number of events is λ, is equal to

$$P(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

For a Poisson distribution, $\mu = \sigma^2 = \lambda$. The mean and variance of any Poisson distribution are both equal to λ.

95% confidence interval for λ in the Poisson distribution

In a Poisson distribution, the mean and variance are both λ. Thus, the standard deviation is $\sigma = \sqrt{\lambda}$. If we can approximate the Poisson distribution by the normal distribution (i.e. if $\lambda > 5$), then we can use $Z_{crit} = 1.96$ to compute the confidence interval. If we observe $k > 10$ Poisson events, then an approximate 95% confidence interval for λ is

$$k \pm 1.96\sqrt{k}.$$

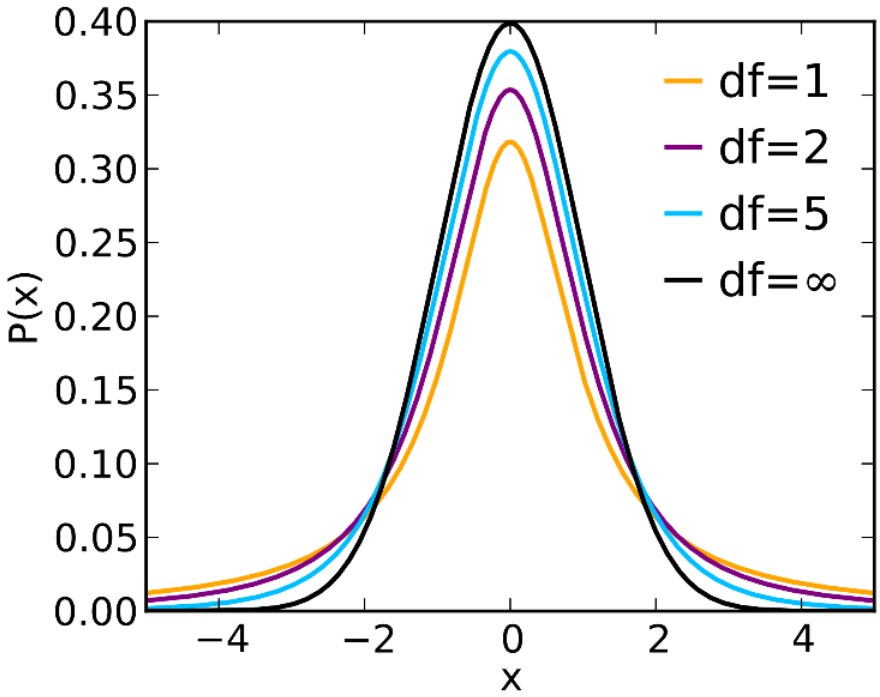
We are 95% confident that λ is within this confidence interval.

What is the variance of the difference between two sample means?

Properties of the variance:

- $Var(ax) = a^2 Var(x)$
- $Var(x + y) = Var(x) + Var(y)$ only if x and y are independent (not correlated)
- $Var(\overline{x_1} - \overline{x_2}) = Var(\overline{x_1} + \text{-}\overline{x_2}) = Var(\overline{x_1}) + Var(\text{-}\overline{x_2})$
- $Var(\overline{x_1}) + (\text{-}1)^2 Var(\overline{x_2}) = Var(\overline{x_1}) + Var(\overline{x_2})$, only if $\overline{x_1}$ and $\overline{x_2}$ are independent.
- If $\overline{x_1}$ and $\overline{x_2}$ are sample means, then their variances are respectively $Var(\overline{x_1}) = \sigma_{x1}^2/n_{x1}$ and $Var(\overline{x_2}) = \sigma_{x2}^2/n_{x2}$.
- So $Var(\overline{x_1} - \overline{x_2}) = \sigma_{x1}^2/n_{x1} + \sigma_{x2}^2/n_{x2}$.

Z versus t



Show a drawing: we have several possible populations, and a sample. Which population did the sample come from? To answer that question we ask "is the sample unusual relative to the reference population¿'

Z test in general

$$Z = \frac{\text{observed statistic} - \text{expected}}{\text{true standard deviation of observed statistic}}$$

If observed statistic is the mean of a sample of n, then its standard deviation is $\sigma/\sqrt{n}$.
If observed statistic is a binomial variate, its standard deviation is $\sqrt{np(1-p)}$.
If observed statistic is a Poisson variate, its standard deviation is $\sqrt{\lambda}$.

Show a drawing of the normal distribution with tails, on the Z scale. This represents the reference population. We can plot our sample on this reference population. if the sample is in one of the tails, then we conclude the sample is unusual and it was not drawn from the reference population.

Z test in general

$$Z = \frac{\text{observed statistic} - \text{expected}}{\text{true standard deviation of observed statistic}}$$

If observed statistic is the mean of a sample of n, then its standard deviation is $\sigma/\sqrt{n}$.
If observed statistic is a binomial variate, its standard deviation is $\sqrt{np(1-p)}$.
If observed statistic is a Poisson variate, its standard deviation is $\sqrt{\lambda}$.

First we have to convert to the Z scale. To do so, we need to know the standard deviation, the denominator of the Z expression. If the observation is the mean of the sample, the standard deviation is the SEM, or the standard deviation of the distribution divided by the square root of the sample size. If the sample size is one, then we have just one observation, and the standard deviation in the denominator is just the standard deviation of the distribution.

Take some examples from Problem sets 5 and 6. The orchard question, question 2 on PS 5. Here we have the sample mean, so we use the SEM in the denominator.

Then Mendel's pea experiment. Here we have the binomial distribution, so we can use the binomial variance to calculate the Z. The denominator is just the square root of the binomial variance. We can assume a normal distribution because $np > 5$.

Then try a Poisson example, with the Poisson we also know what the true standard deviation is, if we know the true mean. Let's say we know the true mutation rate is 1 mutation per seed. We screen 100 seeds and find a total of 118 mutations. Is that unusual if the normal mutation rate is 1 per seed? The Z value would be (118-100)/(sqrt(100)) = 1.8, which is not quite unusual, but close, since our convention is a 2-tail probability of 0.05, and the Z value corresponding to that 2-tail probability is 1.96. We can assume normality because we added up 100 seeds, and the sum should be close to normal according to the Central Limit Theorem.

What is the variance of the difference between two sample means?

Properties of the variance:

- $Var(ax) = a^2 Var(x)$
- $Var(x + y) = Var(x) + Var(y)$ only if x and y are independent (not correlated)
- $Var(\overline{x_1} - \overline{x_2}) = Var(\overline{x_1} + \text{-}\overline{x_2}) = Var(\overline{x_1}) + Var(\text{-}\overline{x_2})$
- $Var(\overline{x_1}) + (\text{-}1)^2 Var(\overline{x_2}) = Var(\overline{x_1}) + Var(\overline{x_2})$, only if $\overline{x_1}$ and $\overline{x_2}$ are independent.
- If $\overline{x_1}$ and $\overline{x_2}$ are sample means, then their variances are respectively $Var(\overline{x_1}) = \sigma^2_{x1}/n_{x1}$ and $Var(\overline{x_2}) = \sigma^2_{x2}/n_{x2}$.
- So $Var(\overline{x_1} - \overline{x_2}) = \sigma^2_{x1}/n_{x1} + \sigma^2_{x2}/n_{x2}$.

We can also use a Z test to test for a difference between two samples. If we have two samples: were they drawn from the same distribution with true mean and standard deviation, or were they drawn from different distributions? Our observed statistic is the difference between the two sample means. Our expected value of the difference is (usually) zero. And in the denominator, the standard deviation of the difference is the square root of the sum of the variances of the two sample means.

Z test in general

$$Z = \frac{\text{observed statistic} - \text{expected}}{\text{true standard deviation of observed statistic}}$$

If observed statistic is the difference between two sample means, $\overline{x_1} - \overline{x_2}$, then the expected is usually zero, and

the standard deviation of the difference is

$$\text{standard deviation of difference} = \sqrt{\sigma^2_{x1}/n_{x1} + \sigma^2_{x2}/n_{x2}}.$$

So

$$Z = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{\sigma^2_{x1}/n_{x1} + \sigma^2_{x2}/n_{x2}}}$$

Up until now we have assumed we know the true value of the standard deviation. We needed to know that because we needed to assume it was a constant, so we could assume that Z was normal. If you divide a normal variate with a constant, then the resulting variate is also normal. But if we don't know the true value of the standard deviation, we can calculate it from the sample. But then, it is no longer a constant, it is a variable. Every time we take a sample, the calculated standard deviation is slightly different. In that case, the distribution of the Z will no longer be normal. It will be more spread out than normal. Because it is no longer normal, it is no longer called a Z variate or Z score or Z distribution. Rather it is called a t distribution. The shape of a t distribution was derived by William Sealey Gossett (1876 - 1937) a British statistician who worked for the Guinness brewery. He called himself "Student", his pen-name, and the t distribution is also called "Student's t distribution".

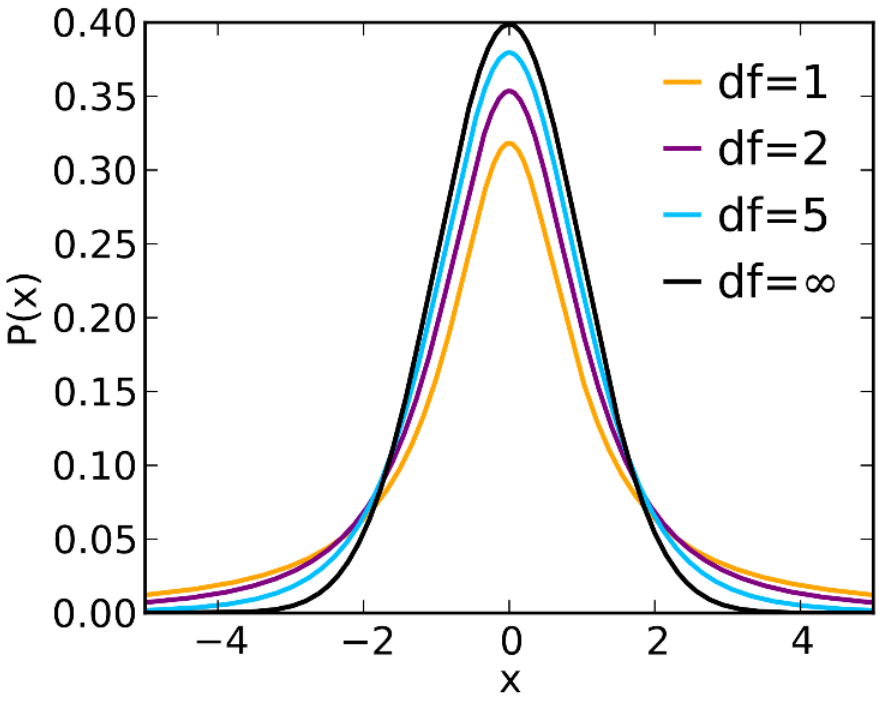
Z versus t

$$Z = \frac{\overline{x} - \text{expected}}{\frac{\sigma}{\sqrt{n}}}$$

$$t = \frac{\overline{x} - \text{expected}}{\frac{s}{\sqrt{n}}}$$

$$s = \text{sample standard deviation} = \sqrt{\frac{\sum_1^n (x - \overline{x})^2}{n-1}}$$

Z versus t



Lecture 12: confidence intervals continued

Announcements:

- Reading for this week's subject: pp. 347 - 353 in Horvat (reading for this week).
- Midterm Exam 1 was postponed to next Tuesday, due to power outage.
- No lab exercises next week.
- Practice using Midterm 1 practice exam on the website.
- Quiz 6 is due the day after the midterm, Wednesday November 23.

Calculating the 95% confidence interval

$$P\left(\bar{x} - 1.96\frac{\sigma}{\sqrt{n}} \leq \mu \leq \bar{x} + 1.96\frac{\sigma}{\sqrt{n}}\right) = 0.95.$$

$$\bar{x} \pm 1.96\frac{\sigma}{\sqrt{n}}$$

This is the 95% confidence interval for the sample mean.

We are 95% confident that the true mean is contained in this interval.

$$Z_{0.025} = 1.96.$$

$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

Calculating the 95% confidence interval from the binomial distribution

$$P\left(\hat{p} - 1.96\sqrt{p(1-p)/n} \leq p \leq \hat{p} + 1.96\sqrt{p(1-p)/n}\right) = 0.95.$$

$$\hat{p} \pm 1.96\sqrt{p(1-p)/n}$$

This is the 95% confidence interval for the true probability of success p.

We are 95% confident that p is contained in this interval.

$$Z_{0.025} = 1.96.$$

$$\bar{x} \pm Z_{0.025}\frac{\sigma}{\sqrt{n}}.$$

Poisson distribution

The Poisson distribution can be thought of a special case of the binomial distribution for very large n and very small p.

In that case, the mean np and variance $np(1-p)$ are essentially equal to each other.

Poisson distribution

The probability of exactly k events during some interval or across some spatial region, given that the mean number of events is λ, is equal to

$$P(k, \lambda) = \frac{\lambda^k e^{-\lambda}}{k!}$$

For a Poisson distribution, $\mu = \sigma^2 = \lambda$. The mean and variance of any Poisson distribution are both equal to λ.

95% confidence interval for λ in the Poisson distribution

In a Poisson distribution, the mean and variance are both λ. Thus, the standard deviation is $\sigma = \sqrt{\lambda}$. If we can approximate the Poisson distribution by the normal distribution (i.e. if $\lambda > 5$), then we can use $Z_{crit} = 1.96$ to compute the confidence interval. If we observe $k > 10$ Poisson events, then an approximate 95% confidence interval for λ is

$$k \pm 1.96\sqrt{k}.$$

We are 95% confident that λ is within this confidence interval.

Confidence intervals can be used to test a hypothesis

Let's say we'd like to buy a batch of apples if we're 95% confident that their mean mass is greater than 200 g.

We randomly sample several apples from the batch and find the 95% confidence interval for μ is (189, 195).

Is this consistent with the proposition that $\mu > 200$?

No.

Is this consistent with the proposition that $\mu < 200$?

Yes.

So we can conclude, with 95% confidence that $\mu < 200$.

Do we buy the apples?

No.

Confidence intervals can be used to test a hypothesis

Let's say we'd like to buy a batch of apples if we're 95% confident that their mean mass is greater than 200 g.

We randomly sample several apples from the batch and find the 95% confidence interval for μ is (198, 205).

Is this consistent with the proposition that $\mu > 200$?

Yes.

Is this consistent with the proposition that $\mu < 200$?

Yes.

Do we buy the apples?

No.

Confidence intervals can be used to test a hypothesis

Let's say we'd like to buy a batch of apples if we're 95% confident that their mean mass is greater than 200 g.

We randomly sample several apples from the batch and find the 95% confidence interval for μ is (205, 210).

Is this consistent with the proposition that $\mu > 200$?

Yes.

Is this consistent with the proposition that $\mu < 200$?

No.

So we can conclude, with 95% confidence that $\mu > 200$.

Do we buy the apples?

Yes.

Confidence intervals can be used to test a hypothesis

Let's say we'd like to buy a batch of apples if we're 95% confident that their mean mass is greater than 200 g.

He find this confidence interval: (205, 210), we have **rejected the hypothesis** that $\mu < 200$. We have rejected the hypothesis at the **95% confidence level**. You can also say we have rejected the hypothesis at the **5% significance level**. The significance level is also called alpha, α.

$$\alpha = 1 - \text{confidence level}$$

.

Example: poll results: binomial confidence interval

Let's say a poll of 1000 random Croatians found that 47.5% were in favor of entering the EU. Do a majority of Croatians favor entering the EU? Get the confidence interval. So $\hat{p} = 0.475$. The 95% confidence interval for the true proportion p is

$$\hat{p} \pm 1.96\sqrt{\hat{p}(1-\hat{p})/n}$$

$$= 0.475 \pm 1.96\sqrt{0.475(1-0.475)/1000}$$

$$= 0.475 \pm 0.03$$

$$= (0.445, 0.505)$$

Is this consistent with a majority favoring entering the EU? Yes. Is this consistent with a majority opposing entering the EU? Yes. We can't conclude anything at the 95% confidence level.

Example: poll results: binomial confidence interval

Let's say a poll of 1000 random Croatians found that 46% were in favor of entering the EU. Do a majority of Croatians favor entering the EU?

The 95% confidence interval for p is (0.43, 0.49).

Is this consistent with a majority favoring entering the EU?

No.

Is this consistent with a majority opposing entering the EU?

Yes.

We conclude that the majority of Croatians are opposed to entering the EU.

This is the 95% confidence level, or the 5% significance level.

Example: poll results: binomial confidence interval

Let's say a poll of 1000 random Croatians found that 54% were in favor of entering the EU. Do a majority of Croatians favor entering the EU?

The 95% confidence interval for p is (0.51, 0.57).

Is this consistent with a majority favoring entering the EU?

Yes.

Is this consistent with a majority opposing entering the EU?

No.

We conclude that the majority of Croatians are in favor of entering the EU.

This is the 95% confidence level, or the 5% significance level.

So far we've been looking at one population. If we want to compare the means of two populations, what do we do? We can see if their 95% confidence intervals overlap. If they don't, then we conclude that the true means are different. If they do overlap, the true means may still be different, we don't know. We can test for a difference directly by calculating the 99% confidence interval of the difference.

This is equal to the observed difference between the sample means, plus or minus 1.96 times the standard deviation of the difference between the sample means. To get the standard deviation of the difference between the sample means, we need to use the properties of the variance:

What is the variance of the difference between two sample means?

Properties of the variance:

- $Var(ax) = a^2 Var(x)$
- $Var(x + y) = Var(x) + Var(y)$ only if x and y are independent (not correlated)
- $Var(\overline{x_1} - \overline{x_2}) = Var(\overline{x_1} + \text{-}\overline{x_2}) = Var(\overline{x_1}) + Var(\text{-}\overline{x_2})$
- $= Var(\overline{x_1}) + (\text{-}1)^2 Var(\overline{x_2}) = Var(\overline{x_1}) + Var(\overline{x_2})$, only if $\overline{x_1}$ and $\overline{x_2}$ are independent.
- If $\overline{x_1}$ and $\overline{x_2}$ are sample means, then their variances are respectively $Var(\overline{x_1}) = \sigma_1^2/n_1$ and $Var(\overline{x_2}) = \sigma_2^2/n_2$.
- So $Var(\overline{x_1} - \overline{x_2}) = \sigma_1^2/n_1 + \sigma_2^2/n_2$.

So the standard deviation of the difference between two sample means must be equal to

$$s_{\text{diff}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \sqrt{SEM_1^2 + SEM_2^2}$$

.

So the 95% confidence interval of the difference between two sample means is

$$\overline{x_1} - \overline{x_2} \pm 1.96\sqrt{SEM_1^2 + SEM_2^2}.$$

Z test in general

$$Z = \frac{\text{observed statistic} - \text{expected}}{\text{true standard deviation of observed statistic}}$$

If observed statistic is the mean of a sample of n, then its standard deviation is $\sigma/\sqrt{n}$.
If observed statistic is a binomial variate, its standard deviation is $\sqrt{np(1-p)}$.
If observed statistic is a Poisson variate, its standard deviation is $\sqrt{\lambda}$.
If observed statistic is the difference between two sample means, its standard deviation is $\sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2}$.

Example: poll results Z test

Let's say a poll of 1000 random Croatians found that 47.5% Were in favor of entering the EU. Is this unusual if the true proportion was 51.0% in favor?

The Z value for this calculation is -35/15.80823 = -2.214037. The associated probability is 0.01341312, so this is an unusual result, and we conclude that Croatians as a whole are not in favor of entering the EU according to this poll.

Chi square test

$$\chi^2 = \frac{\sum_1^n (\text{count of observed} - \text{count of expected})^2}{\text{count of expected}}$$

Degrees of freedom is the number of counts that are free to vary in this calculation.

Example: poll results chi square test

Let's say a poll of 1000 random Croatians found that 47.5% Were in favor of entering the EU. Is this unusual if the true proportion was 51.0% in favor?

We can redo this problem as a chi-square test, which is mathematically equivalent to a Z test. The chi square statistic is $\chi^2 = 35^2/510 + 35^2/490 = 4.901961 = -2.214037^2 = Z^2$. So a question involving a binomial variable can be answered either by a Z test or a chi-square test, assuming that the normality assumption of the Z test is met, namely that the binomial variate is normally distributed, which is generally assumed to be true if the mean of the binomial variable is greater than 5 ($np(1-p) > 5$).

If you have a choice between a Z test and a chi-square test, which should you choose? Because they are mathematically identical, it makes little difference. However, with the Z test it is possible for the Z value to be positive or negative. It is not possible for the chi-square value to be negative. So if you use a chi-square test, you need to keep in mind the direction of the difference. In the example above 475 Croatians were in favor and 525 were opposed to EU entry. If these numbers were reversed, they would have given the same chi-square statistic. While the chi-square test is insensitive to the direction of any result, the Z test is sensitive.

Z test versus t test

What if we don't know σ? We calculate it from the sample as the sample standard deviation:

$$s = \sqrt{\frac{\sum_1^n (x_i - \overline{x})^2}{n-1}}$$

$n-1$ is called v, or the "degrees of freedom" used in the calculation of s.

Z test versus t test

$$Z = \frac{\text{observed} - \text{expected}}{\text{true standard deviation of observed}}$$

$$Z = \frac{\text{observed} - \text{expected}}{\sigma_{\text{obs}}}.$$

$$t = \frac{\text{observed} - \text{expected}}{\text{estimated standard deviation of observed}}$$

$$t = \frac{\text{observed} - \text{expected}}{s_{\text{obs}}}.$$

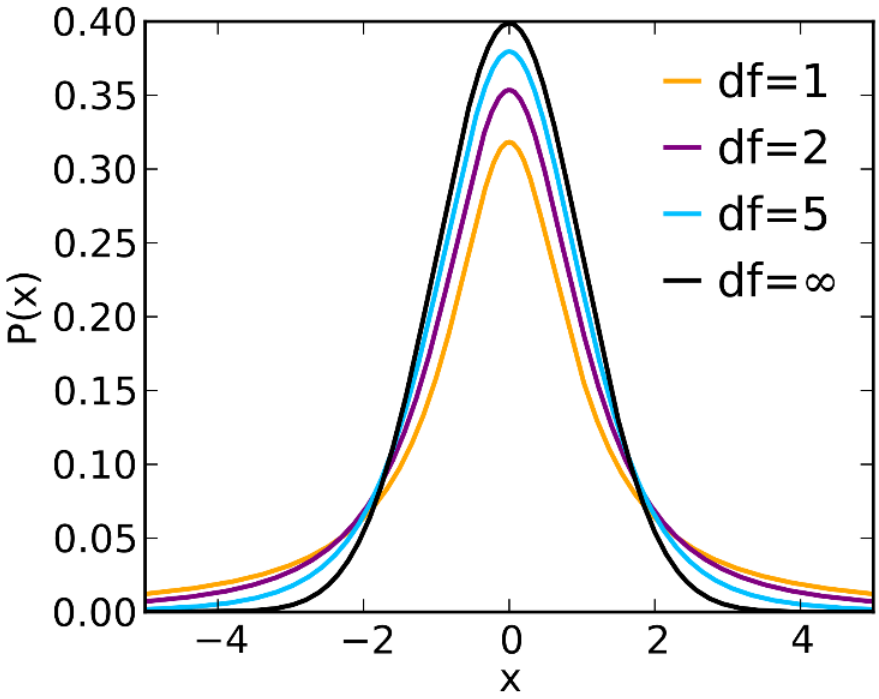
Z test versus t test

Z is normally distributed, with mean of 0 and standard deviation of 1.

t is very close to normally distributed, symmetrical, with mean of 0 and standard deviation slightly higher than 1.

But t is practically indistinguishable from a normal distribution if $n > 30$.

t distribution for various sample sizes (or degrees of freedom)



Example: a sample of heights

Say we have a random sample of 15 women, their sample mean height is $\overline{x} = 173$ cm, and sample standard deviation of height is $s = 9.2$ cm. Is this unusual if the population mean is $\mu = 169$ cm?

Example: exam scores

In our Exam 1, the mean score of women was 72.7 and the mean score for men was 70.3. The standard deviation for women was 7.7 and for men was 8.3. There were 14 women and 13 men. Is the sex difference in means unusual, if we expect the difference to be zero?

12 Hypothesis testing

Lecture 13: Hypothesis testing

Announcements:

- Reading for this week's subject: pp. 61 - 79 in Vasilj.
- Midterm Exam 1 is tomorrow during lab session, SK-INF.

- Rules: 50 minutes, closed book, closed internet, open one page of notes, your choice.
- Quiz 6 due Wednesday, day after exam.

Testing a hypothesis with a confidence interval

Let's say we plant 1000 asparagus seeds and found that 46% were female. Is this consistent with a 1:1 sex ratio in asparagus, i.e. 50% female and 50% male? The 95% confidence interval for p is (0.43, 0.49). Is this consistent with a true sex ratio of 1:1? No. Is this consistent with a true sex ratio different from 1:1? Yes. We conclude that the sex ratio in asparagus is not 1:1. We reject the hypothesis that the sex ratio in asparagus is 1:1. This is the 95% confidence level, or the 5% significance level.

Last time we learned how to test a hypothesis by calculating the confidence interval. If the hypothesis lies outside the interval, then the hypothesis is rejected. We say we reject it at the 95% level, if we calculated the 95% confidence interval.

So far we've been looking at one population. If we want to compare the means of two populations, what do we do? We can see if their 95% confidence intervals overlap. If they don't, then we conclude that the true means are different. If they do overlap, the true means may still be different, we don't know. We can test for a difference directly by calculating the 95% confidence interval of the difference.

This is equal to the observed difference between the sample means, plus or minus 1.96 times the standard deviation of the difference between the sample means. To get the standard deviation of the difference between the sample means, we need to use the properties of the variance:

What is the variance of the difference between two sample means?

Properties of the variance:

- $Var(ax) = a^2 Var(x)$
- $Var(x + y) = Var(x) + Var(y)$ only if x and y are independent (not correlated)
- $Var(\overline{x_1} - \overline{x_2}) = Var(\overline{x_1} + \text{-}\overline{x_2}) = Var(\overline{x_1}) + Var(\text{-}\overline{x_2})$
- $= Var(\overline{x_1}) + (\text{-}1)^2 Var(\overline{x_2}) = Var(\overline{x_1}) + Var(\overline{x_2})$, only if $\overline{x_1}$ and $\overline{x_2}$ are independent.
- If $\overline{x_1}$ and $\overline{x_2}$ are sample means, then their variances are respectively $Var(\overline{x_1}) = \sigma_1^2/n_1$ and $Var(\overline{x_2}) = \sigma_2^2/n_2$.
- So $Var(\overline{x_1} - \overline{x_2}) = \sigma_1^2/n_1 + \sigma_2^2/n_2$.

Standard deviation of the difference between two sample means

So the standard deviation of the difference between two sample means must be equal to

$$s_{\text{diff}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \sqrt{SEM_1^2 + SEM_2^2}$$

.

So the 95% confidence interval of the difference between two sample means is

$$\overline{x_1} - \overline{x_2} \pm 1.96\sqrt{SEM_1^2 + SEM_2^2}.$$

Example

Let's say we have a new genotype of apple, and we want to know if its mean mass is the same as the old genotype. We sample 100 apples of each genotype, and found the mean mass of the new genotype is 200 g. The mean mass of the old genotype is 190 g. The standard deviation in mass of both genotypes is 50 g. What is the 95% confidence interval of the difference? What is our margin of error for estimating the difference?

See the tablet for this lecture for the above calculations. Basically, after we calculate the 95% confidence interval for the difference between the two sample means, we just ask: does the confidence interval contain 0? If it does, then we can't reject the hypothesis that the two means are equal (i.e. their difference is 0). If the confidence interval does not contain zero, then we conclude that the two population means are different, i.e. that one of the means is greater than the other.

Hypothesis testing: more elegant procedure than calculating the CI

Let's say we hypothesize that the mean apple mass is μ_0. Can we reject this hypothesis? Calculate the 95% confidence interval, call it (L, U) where L is the lower limit of the confidence interval, and U is the upper limit. There are two ways to reject the hypothesis: If $U < \mu_0$ then we can reject the hypothesis. If $L > \mu_0$ then we can reject the hypothesis.

Hypothesis testing: more elegant procedure than calculating the CI

Let's start by considering that $U < \mu_0$ then we can reject the hypothesis. But $U = \bar{x} + 1.96(\text{SEM})$. Then we reject the hypothesis when $\bar{x} + 1.96(\text{SEM}) < \mu_0$. This implies that we reject the hypothesis when $\bar{x} - \mu_0 < -1.96(\text{SEM})$. Or, we reject when

$$\frac{\bar{x} - \mu_0}{\text{SEM}} < -1.96.$$

But, we already know that $Z_{\text{observed}} = \frac{\bar{x} - \mu_0}{\text{SEM}}$. So we can conclude that we reject when

$$Z_{\text{obs}} < -1.96$$

. If this is true, then the confidence interval is entirely below μ_0.

Hypothesis testing: more elegant procedure than calculating the CI

Now let's look at the other criterion for rejecting the hypothesis: If $L > \mu_0$ then we can reject the hypothesis. But $L = \bar{x} - 1.96(\text{SEM})$. Then we reject the hypothesis when $\bar{x} - 1.96(\text{SEM}) > \mu_0$. This implies that we reject the hypothesis when $\bar{x} - \mu_0 > 1.96(\text{SEM})$. Or, we reject when

$$\frac{\bar{x} - \mu_0}{\text{SEM}} > 1.96.$$

But, we already know that $Z_{\text{observed}} = \frac{\bar{x} - \mu_0}{\text{SEM}}$. So we can conclude that we reject when

$$Z_{\text{obs}} > 1.96.$$

If this is true, then the confidence interval is entirely above μ_0.

Hypothesis testing: more elegant procedure than calculating the CI

So, the more elegant procedure for hypothesis testing is to just calculate the Z value of our observed mean, or difference between two means. Then ask, is

$$|Z_{\text{observed}}| > 1.96$$
$$\left|\frac{\bar{x} - \mu_0}{\text{SEM}}\right| > 1.96$$

If this condition is met, then we reject our hypothesis.

If $|Z_{\text{observed}}| > 1.96$, then this means that the 95% confidence interval is entirely greater than the hypothesis, and it is therefore rejected at the 95% confidence level. If $|Z_{\text{observed}}| < -1.96$, then the entire 95% confidence interval is less than the hypothesis, so again we reject the hypothesis at the 95% level.

Example

I flip a coin 100 times and get 46 heads, so $\hat{p} = 0.46$. Is this a fair coin? In other words, does $p = 0.5$? We calculate the observed Z and ask: Is $|Z| > 1.96$?

$$Z_{\text{observed}} = \frac{\hat{p} - \mu_0}{\sqrt{\hat{p}(1-\hat{p})/n}}$$
$$Z_{\text{observed}} = \frac{0.46 - 0.5}{\sqrt{0.46(1-0.46)/100}}$$
$$Z_{\text{observed}} = -\frac{0.040}{0.050}$$
$$Z_{\text{observed}} = -0.8.$$

Since $|Z_{\text{observed}}| = 0.8 < 1.96$, we conclude that we can't reject the hypothesis that the coin is fair. This is the 95% confidence interval, or the 5% significance level.

Z test and chi square tests are mathematically equivalent

I flip a coin 100 times and get 46 heads, so $\hat{p} = 0.46$. Is this a fair coin? In other words, does $p = 0.5$? We tested this by calculating $Z_{\text{observed}} = 0.8$. We can also test this by calculating χ^2, the chi-square value of our observation.

$$\chi^2_{\text{observed}} = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$\chi^2_{\text{observed}} = \frac{(46-50)^2 + (54-50)^2}{50}$$

$$\chi^2_{\text{observed}} = \frac{32}{50}$$

$$\chi^2_{\text{observed}} = 0.64 = 0.8^2 = Z^2_{\text{observed}}.$$

Since $\chi^2_{\text{observed}} < 1.96^2 = 3.84$, we conclude that we can't reject the hypothesis that the coin is fair. This is the 95% confidence interval, or the 5% significance level.

Lecture 13: Hypothesis testing

- We have a question about a population.
- We take a sample, calculate a statistic.
- We plot that statistic on the expected probability distribution.
- If the statistic is within the tail, we conclude that the sample is "unusual" relative to the reference population.

Hypothesis testing: procedure

1. We ask a yes/no question about a **population**.
2. We answer the question yes, and answer the question no, using symbols for the population means.
3. We label one answer the **null hypothesis** and the other answer the **alternative hypothesis**.
4. We decide the criterion for rejecting the null hypothesis.
5. The test is one of: **two-tailed**, **right-tailed**, or **left-tailed**.
6. We take a sample, and calculate our **test statistic** (Z or t or χ^2 for now)
7. We find if the observed test statistic is in the rejection region (**critical region** or **tail**) of the distribution.
8. If the statistic is in the rejection region, we **reject** the null hypothesis and **accept** the alternative hypothesis.
9. If the statistic is not in the rejection region, we **retain** the null hypothesis, and do not accept the alternative hypothesis.

Examples

I flip a coin n times and get r heads. Is this a fair coin?

First we define the true means of the population we're interested in. The population is all possible coin flips of this particular coin. The population mean could be defined as p, the true proportion heads in an infinite number of flips. So first we answer the question "yes": yes, it is a fair coin would be equivalent to writing the equation $p = 0.5$. If the true proportion heads in an infinite number of flips is 0.5, then the coin is fair. Then we answer the question "no": if the coin is not fair, then the probability of landing heads is something besides 0.5. So the equation for this is $p \neq 0.5$. Now we decide: which of these equations is the null hypothesis and which is the alternative hypothesis? The null hypothesis is the hypothesis containing the equals sign, which is the first equation above. So the null hypothesis H_0 is $p = 0.5$. So

we can write $H_0 : p = 0.5$. This means the other equation is the alternative hypothesis: $H_A : p \neq 0.5$. Now we decide whether this is a 2-tailed, left-tailed, or right-tailed hypothesis. Because the alternative hypothesis contains the $\neq$ sign, we know that this is a two-tailed test. This means the null hypothesis can be rejected in either direction, right or left. To the right, the coin's chances of landing heads is greater than 0.5. To the left, its chances of landing heads are less than 0.5. Next we write down the equation for the statistic that tests the null hypothesis. We have been using either t or Z so far. In this case we can use a Z test if n is sufficiently large (then we have a normal distribution). $Z = \frac{r/n - 0.5}{\sqrt{r/n(1-r/n)/n}}$. What is the criterion for rejection of the null hypothesis? The criterion is $|Z| > Z_{crit}$, where Z_{crit} is the value of Z that defines what the tail is of the distribution. Since this is a two tailed test, and the tails by definition contain α, and α is conventionally set to 0.05, then the Z_{crit} must be the Z value that cuts off a tail of 0.025. This Z value is 1.96. So $|Z| > 1.96$ is the criterion for rejecting the null hypothesis that the coin is fair.

See the solutions to the examples below in the tablet file for this lecture.

Examples

I create a new genotype of olive tree. Do its olives produce more oil than the old genotype?

This is a right-tail test, and the criterion for rejection is If $Z_{\text{observed}} > 1.65$. Note we don't have the absolute value of Z, just Z. The reason we don't use 1.96 is that the rejection area of the curve is always equal to alpha, and if the right tail is the rejection region, then the right tail is equal to alpha. If alpha is 0.05, then we want the Z value that cuts off a right tail equal to 0.05 of the distribution. That Z is 1.65. If it were a left tail test, then the criterion would be $Z_{\text{observed}} < -1.65$.

Rules

- The hypothesis containing the = sign is the null hypothesis, denoted H_0. The other hypothesis is the alternative hypothesis, denoted H_A.
- If the alternative hypothesis contains the $\neq$ sign, the test is two-tailed.
- If the alternative hypothesis contains the $<$ sign, the test is left-tailed.
- If the alternative hypothesis contains the $>$ sign, the test is right-tailed.

Examples

I have a new chemical pesticide. Does it leave less residue on apples than the old kind?

This is a one-tailed test, left-tailed, and the criterion for rejection of the null hypothesis is $Z_{\text{observed}} < -1.64$,

Examples

I take a poll. Are Croatians in favor of entering the EU?

This is a right-tailed test, with criterion for rejection $Z_{\text{observed}} > 1.65$.

Examples

Is the sex ratio of males to females 1:1 in some population?

This is a two-tailed test with criterion for rejection $|Z_{\text{observed}}| > 1.96$.

Here is some terminology for writing the critical values of a test statistic. The subscript is basically alpha divided by the number of tails in the test. I've written t values here, but the same notation applies to any statistic, whether t, Z, χ^2, etc.

Terminology

- The **critical value** of a test statistic is the value that defines the rejection region of the distribution.
- $t_{0.05/2}$ is the critical t-value for $\alpha = 0.05$ and a 2-tailed test. Each tail here contains 0.025 of the probability.

- $t_{0.05/1}$ is the critical t-value for $\alpha = 0.05$ and a 1-tailed test. The single tail contains 0.05 of the probability.
- $t_{0.01/1}$ is the critical t-value for $\alpha = 0.01$ and a 1-tailed test. The single tail contains 0.01 of the probability.

Examples

Is the level of arsenic in a well greater than some minimum value x?
This is a right-tailed test with criterion for rejection $Z_{\text{observed}} > 1.65$.

Lecture 14: Hypothesis testing, continued

Announcements:

- Reading for this week's subject: pp. 61 - 79 in Vasilj.
- Your total course grade is now shown on Moodle in the Grades (Ocjene) window.
- This is your approximate grade, it does not take into account your attendance.
- This is your grade only if you have good attendance.
- This also does not take into account your score on English versions of quizzes. This is because the English versions are not required. However, if you do both English and Croatian versions, I will take this into consideration as extra credit at the end of the class.
- Also shown is your rank in the class, from 1 to 48.

Testing a hypothesis with a confidence interval

Let's say we plant 1000 asparagus seeds and found that 46% were female. Is this consistent with a 1:1 sex ratio in asparagus, i.e. 50% female and 50% male? The 95% confidence interval for p is (0.43, 0.49). Is this consistent with a true sex ratio of 1:1? No. Is this consistent with a true sex ratio different from 1:1? Yes. We conclude that the sex ratio in asparagus is not 1:1. We reject the hypothesis that the sex ratio in asparagus is 1:1. This is the 95% confidence level, or the 5% significance level.

Hypothesis testing: more elegant procedure than calculating the CI

Let's say we hypothesize that the mean apple mass is μ_0. Can we reject this hypothesis? Calculate the 95% confidence interval, call it (L, U) where L is the lower limit of the confidence interval, and U is the upper limit. There are two ways to reject the hypothesis: If $U < \mu_0$ then we can reject the hypothesis. If $L > \mu_0$ then we can reject the hypothesis.

Hypothesis testing: more elegant procedure than calculating the CI

Now let's look at the other criterion for rejecting the hypothesis: If $L > \mu_0$ then we can reject the hypothesis. But $L = \bar{x} - 1.96(\text{SEM})$. Then we reject the hypothesis when $\bar{x} - 1.96(\text{SEM}) > \mu_0$. This implies that we reject the hypothesis when $\bar{x} - \mu_0 > 1.96(\text{SEM})$. Or, we reject when

$$\frac{\bar{x} - \mu_0}{\text{SEM}} > 1.96.$$

But, we already know that $Z_{\text{observed}} = \frac{\bar{x} - \mu_0}{\text{SEM}}$. So we can conclude that we reject when

$$Z_{\text{obs}} > 1.96.$$

If this is true, then the confidence interval is entirely above μ_0.

Hypothesis testing: more elegant procedure than calculating the CI

So, the more elegant procedure for hypothesis testing is to just calculate the Z value of our observed mean, or difference between two means. Then for a two-tailed test, here is our criterion:

$$|\frac{\bar{x} - \mu_0}{\text{SEM}}| > 1.96$$

$$|Z_{\text{observed}}| > 1.96$$

If this condition is met, then we reject our hypothesis.

For a left-tailed test,

$$Z_{\text{observed}} < -1.64$$

For a right-tailed test,

$$Z_{\text{observed}} > 1.64$$

Hypothesis testing: procedure

1. We ask a yes/no question about a **population**.
2. We answer the question yes, and answer the question no, using symbols for the population means.
3. We label one answer the **null hypothesis** and the other answer the **alternative hypothesis**.
4. We decide the criterion for rejecting the null hypothesis.
5. The test is one of: **two-tailed**, **right-tailed**, or **left-tailed**.
6. We take a sample, and calculate our **test statistic** (Z or t or χ^2 for now)
7. We find if the observed test statistic is in the rejection region (**critical region** or **tail**) of the distribution.
8. If the statistic is in the rejection region, we **reject** the null hypothesis and **accept** the alternative hypothesis.
9. If the statistic is not in the rejection region, we **retain** the null hypothesis, and do not accept the alternative hypothesis (yet).

Rules

- If the alternative hypothesis contains the $\neq$ sign, the test is two-tailed.
- If the alternative hypothesis contains the $<$ sign, the test is left-tailed.
- If the alternative hypothesis contains the $>$ sign, the test is right-tailed.

Hypothesis testing: example

We have a new variety of tomato. Does it have **the same** resistance to fusarium wilt as the old variety? To answer this question, we do an experiment. We expose 100 new plants to fusarium, and found that the mean growth rate was 35 with standard deviation 10. We know from previous experience that the mean growth rate of the old genotype is 33.

Examples

I have a new chemical pesticide. Does it leave less residue on apples than the old kind?

This is a one-tailed test, left-tailed, and the criterion for rejection of the null hypothesis is $Z_{\text{observed}} < -1.64$,

Examples

I take a poll. Are Croatians in favor of entering the EU?

This is a right-tailed test, with criterion for rejection $Z_{\text{observed}} > 1.65$.

Examples

Is the sex ratio of males to females 1:1 in some population?

This is a two-tailed test, with criterion for rejection $|Z_{\text{observed}}| > 1.96$.

The null distribution and alternative distribution

Null distribution: the distribution of a variable if the null hypothesis is true Alternative distribution: the distribution of a variable if the alternative hypothesis is true.

Hypothesis testing: example

We have a new variety of tomato. Does it have **the same** resistance to fusarium wilt as the old variety? To answer this question, we do an experiment. We expose 100 new plants to fusarium, and found that the mean growth rate was 35 with standard deviation 10. We know from previous experience that the mean growth rate of the old genotype is 33.

Hypothesis testing: example

We have a new variety of tomato. Does it have **lower** resistance to fusarium wilt than the old variety? To answer this question, we do an experiment. We expose 100 new plants to fusarium, and found that the mean growth rate was 35 with standard deviation 10. We know from previous experience that the mean growth rate of the old genotype is 33.

Hypothesis testing: example

We have a new variety of tomato. Does it have **higher** resistance to fusarium wilt than the old variety? To answer this question, we do an experiment. We expose 100 new plants to fusarium, and found that the mean growth rate was 35 with standard deviation 10. We know from previous experience that the mean growth rate of the old genotype is 33.

Some terminology

- If we reject the null hypothesis, then we conclude our data are **significant**, or **significantly different** from the null.
- The **probability** of the test is the probability of a test result that is of equal or greater departure from the null hypothesis, if the null hypothesis were true.
- In a two-tailed test, the probability is twice the tail probability. In a one-tailed test, the probability is the one-tail probability.
- In R, you can calculate the probability of a Z test with pnorm(). For a two tailed test, the probability of Z_{observed} is 2*(1 - pnorm(abs(Z_{observed}))). This gives the sum of the two tails.
- The probability for a left-tailed test is the left tail. In R the left tail is pnorm(Z_{observed}).
- The probability for a right-tailed test is the right tail. In R the right tail is 1 - pnorm(Z_{observed}).

13 Statistical power

Lecture 15: Type I, II errors, power of a hypothesis test

Announcements:

- Quiz 7 is up and due on Wednesday.
- Overall course grades are now posted (ignoring attendance).
- If you have questions, stay after the first hour.
- On the menu today, examples of t tests and statistical power.

Some terminology

- If we reject the null hypothesis, then we conclude our data are **significant**, or **significantly different** from the null.
- The **probability** of the test is the probability of a test result that is of equal or greater departure from the null hypothesis, if the null hypothesis were true.
- In a two-tailed test, the probability is twice the tail probability. In a one-tailed test, the probability is the one-tail probability.
- In R, you can calculate the probability of a Z test with pnorm(). For a two tailed test, the probability of Z_{observed} is 2*(1 - pnorm(abs(Z_{observed}))). This gives the sum of the two tails.
- The probability for a left-tailed test is the left tail. In R the left tail is pnorm(Z_{observed}).
- The probability for a right-tailed test is the right tail. In R the right tail is 1 - pnorm(Z_{observed}).

Z test in general

$$Z = \frac{\text{observed statistic} - \text{expected}}{\text{true standard deviation of observed statistic}}$$

If observed statistic is the mean of a sample of n, then its standard deviation is $\sigma/\sqrt{n}$, the SEM.
If observed statistic is a binomial proportion, its standard deviation is $\sqrt{p(1-p)/n}$.
If observed statistic is a Poisson variate, its standard deviation is $\sqrt{\lambda}$.
If the observed statistic is the difference between two sample means, its standard deviation is $\sqrt{SEM_1^2 + SEM_2^2}$.

Standard deviation of the difference between two sample means

So the standard deviation of the difference between two sample means must be equal to

$$s_{\text{diff}} = \sqrt{\sigma_1^2/n_1 + \sigma_2^2/n_2} = \sqrt{SEM_1^2 + SEM_2^2}$$

.

So the 95% confidence interval of the difference between two sample means is

$$\overline{x_1} - \overline{x_2} \pm 1.96\sqrt{SEM_1^2 + SEM_2^2}.$$

Z versus t

$$Z = \frac{\overline{x} - \text{expected}}{\frac{\sigma}{\sqrt{n}}}$$

$$t = \frac{\overline{x} - \text{expected}}{\frac{s}{\sqrt{n}}}$$

$$s = \text{sample standard deviation} = \sqrt{\frac{\sum_1^n (x - \overline{x})^2}{n-1}}$$

the "degrees of freedom", or "df" is $n - 1$.

Example

Let's say we have a new genotype of apple, and we want to know if its mean mass is the same as the old genotype. We sample 100 apples of each genotype, and found the mean mass of the new genotype is 200 g. The mean mass of the old genotype is 190 g. The standard deviation in mass of both genotypes is 50 g. What is the 95% confidence interval of the difference? What is our margin of error for estimating the difference? What is the Z_{observed} for the difference? Do we reject the null hypothesis?

Example

Let's say we have a new genotype of apple, and we want to know if its mean mass is the same as the old genotype. We sample 10 apples of each genotype, and found the mean mass of the new genotype is 200 g. The mean mass of the old genotype is 190 g. The standard deviation in mass of both genotypes is 50 g. What is the 95% confidence interval of the difference? What is our margin of error for estimating the difference? What is the t_{observed} for the difference? Do we reject the null hypothesis?

Problem: statistical power

Statistical significance is not the same as biological "significance" (or meaning).

If you do not reject the null hypothesis, that could be because:
1) the null hypothesis is true, or
2) your test was not powerful enough: your sample size was too low.

What is the power of a test?

Let's say you do a t-test to compare the mean weight of apples from grower 1 to grower 2, and find that you can't reject the null hypothesis, and there is no "significant" difference. Does that mean there is really no true difference in weight of apples between the two growers? No. There could be a difference, but your test was not sensitive enough, i.e. its power was too low. How do you increase power? By increasing the sample size. If you increase the sample size, your estimate of the mean apple weight gets better and better, because the standard error of the mean gets smaller and smaller, so eventually, with high enough sample size, you can now reject the null hypothesis.

So this begs the question: what sample size do you need to increase the power of the test enough to reject the null hypothesis?

Problem

As you recall, a screening test has four combinations of test outcome/condition.

In a screening test, a false negative occurs because power is not 100%. If you performed more than one test on a person, you might find that the first negative was a fluke, and that subsequent tests were positive. This might reduce the probability that you will falsely conclude that there is no condition (or disease etc.). So increasing sample size reduces the probability of false negative conclusions, and therefore increases the power of a test.

How do you calculate the power of a hypothesis test?

Example

As you recall, a screening test has four combinations of test outcome/condition.

Errors

Situation	Decision	
	Accept H_0	Reject H_0
H_0 true	$1-\alpha$	α
H_0 false	β	$1-\beta$

Terminology

- **Type I error** is α.
- **Type II error** is β.

- **Sensitivity** of the test is the power, which is $1 - \beta$.
- **Specificity** of the test is $1 - \alpha$.

Basic power equations, one-sample

$$Z$$

-test

$$Z_{pow} = |Z_{crit}| - \frac{|\delta|}{\frac{\sigma}{\sqrt{n}}} = |Z_{crit}| - \frac{|\delta|\sqrt{n}}{\sigma}$$

$$n = \sigma^2 \frac{(Z_{pow} - Z_{crit})^2}{\delta^2}$$

$$|\delta| = \sigma \frac{(Z_{crit} - Z_{pow})}{\sqrt{n}}$$

Calculating power using R

$$\text{Power} = 1 - \text{pnorm}(Z_{pow})$$

$$Z_{pow} = \text{qnorm}(1 \ - \ \text{power})$$

$$Z_{crit} = \pm\text{qnorm}(1 \ - \alpha/T),$$

T is the number of tails in the test. Note that the Z_{crit} is negative for a left-tailed test.

Also, the function power.examp() in TeachingDemos is a nice graphical demonstration of power. For example, you type power.examp(diff = 2, alpha = 0.05, stdev = 10, n = 20) to compute the power of detecting a true difference of 2, at $\alpha = 0.05$, for a standard deviation of 10 and sample size of 20.

Example

A buyer wants to buy a large batch of apples from a grower. But he does not want to buy if the true mean weight is less than 150 g. He can't weigh all the apples, but he can weigh a sample of n, and calculate the sample mean. Let's say the true standard deviation is 30g. If he weighs 20 apples and does a Z test he may or may not reject the null hypothesis. If the **true** mean weight of apples is 150 g, what is the probability that he rejects the null hypothesis, and rejects the apples? Assume he is doing a one-tailed Z-test, with a significance level of 0.05 ($\alpha = 0.05$). He rejects the apples only if he rejects the null hypothesis.

Example

What if the true mean weight of the grower's apples is 149 g? Will he reject the null hypothesis? 145 g? 140 g?

Example

What is the least significant difference (LSD) that the buyer must see before he rejects the grower's apples?

Example

What is the probability that he will see the LSD if the true mean weight of the grower's apples is 149 g? 145 g? 140 g?

Power goes up with sample size.

Example

If his sample size has to be 20, then what true difference (δ) will he be able to detect with his test 80% of the time? What if the sample size is 30? 50? 100? 500?

Example

If he wants to detect, at a power of 80%, a true difference of only 2 g, what sample size will he need? 1 g? 0.5 g?

You need much larger sample sizes to detect smaller true differences.

Example

If he wants to detect a true difference of 10 g, at a power of 80%, what should his sample size be? At a power of 90%? 95%? 99%?

You can detect smaller true differences with higher sample sizes.

Example

If he wants to detect a true difference of 15 g, at a power of 80%, with a sample size of 20. What effect α should he use?

If α is smaller, power also decreases.

14 Testing for equality of variances: the F-test

Lecture 16: Testing for equality of variances: the F-test

Announcements:

- Quiz 8 is posted and due Wednesday, December 7.

Power tradeoffs

- We can use power.examp() in TeachingDemos to visualize:
- Power increases if alpha increases
- Power increases with increasing δ
- Power increases with decreasing σ
- Power increases with increasing n

Basic power equations, one-sample

$$Z$$

-test

$$Z_{pow} = |Z_{crit}| - \frac{|\delta|}{\frac{\sigma}{\sqrt{n}}} = |Z_{crit}| - \frac{|\delta|\sqrt{n}}{\sigma}$$

$$n = \sigma^2 \frac{(Z_{pow} - Z_{crit})^2}{\delta^2}$$

$$|\delta| = \sigma \frac{(Z_{crit} - Z_{pow})}{\sqrt{n}}$$

Calculating power using R

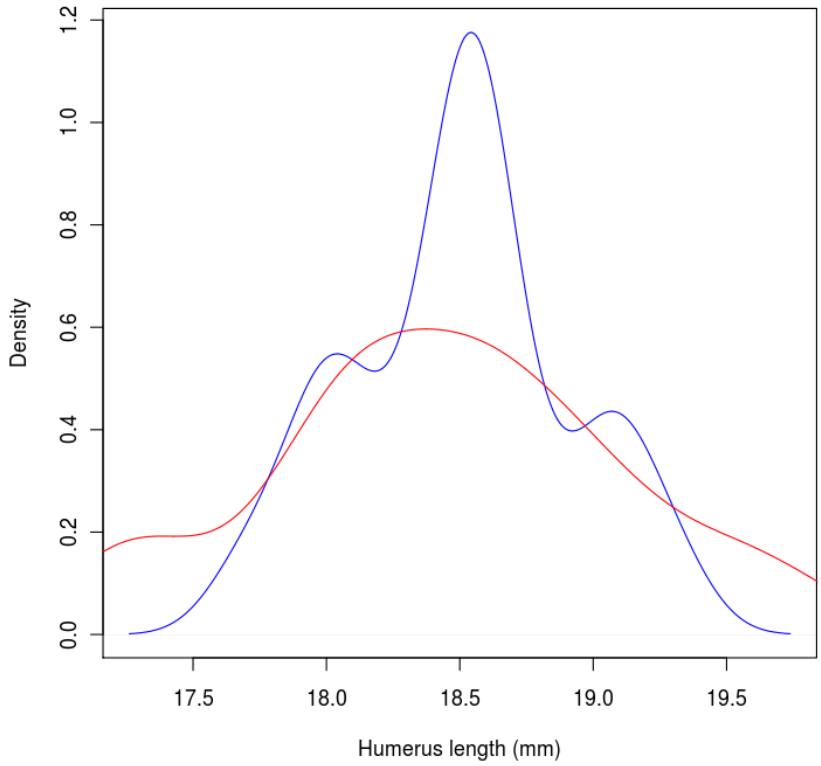
$$\text{Power} = 1 - \text{pnorm}(Z_{pow})$$
$$Z_{pow} = \text{qnorm}(1 - \text{power})$$
$$Z_{crit} = \pm\text{qnorm}(1 - \alpha/T),$$

T is the number of tails in the test. Note that the Z_{crit} is negative for a left-tailed test.

Up until now, we have been using two samples for testing of the equality of the means of two populations. Another common question in biology is whether the variances of two populations are the same. Remember variance is a measure of the spread of a distribution, and the standard deviation is the square root of the variance.

Remember how variance is defined: differently for a continuous variable and a discrete variable.

F-test purpose: to compare two variances



F-test purpose: to compare two variances

F-test purpose: to compare two variances

Measures of variability

- Variance $= Var(x) = \sigma_x^2$
- Standard deviation $= \sqrt{Var(x)} = \sigma_x$

Variance of a distribution

Definition 48. The variance of a discrete probability distribution is

$$Var(x) = E(x-\mu)^2 = \sum_i (x_i - \mu)^2 P(x_i).$$

The variance of a continuous probability distribution is

$$Var(x) = E(x-\mu)^2 = \int_{-\infty}^{+\infty} (x-\mu)^2 f(x) dx.$$

These are the definitions of the absolute variance from a probability distribution.

We can visualize changes in the variance of a distribution using R, by plotting the distribution. For a normal distribution, we can do a plot with dnorm(). For example, set the xvalues as xval = seq(-3, 3, .01). Then plot with plot(xval, dnorm(xval), type = "l"). Then you can see how changes in the variance affect changes in the distribution with lines(xval, dnorm(xval, sd = 0.8), type = "l").

Now we imagine that we take a random sample from that distribution, and estimate the true variance from the sample.

Calculation of the sample variance

$$\text{Sample variance} = s^2 = \frac{\sum_1^n (x_i - \overline{x})^2}{n-1}$$

$$\text{Sample standard deviation} = s = \sqrt{s^2}$$

$n - 1$ is called v, or the "degrees of freedom" used in the calculation of s.

We can calculate variance of a sample x in R with var(x). We can take a random sample from a normal distribution with the function rnorm(). So for example x = rnorm(100) creates 100 datapoints from a normal distribution with mean 0 and variance 1. We can calculate the sample variance of these data with var(rnorm(100)).

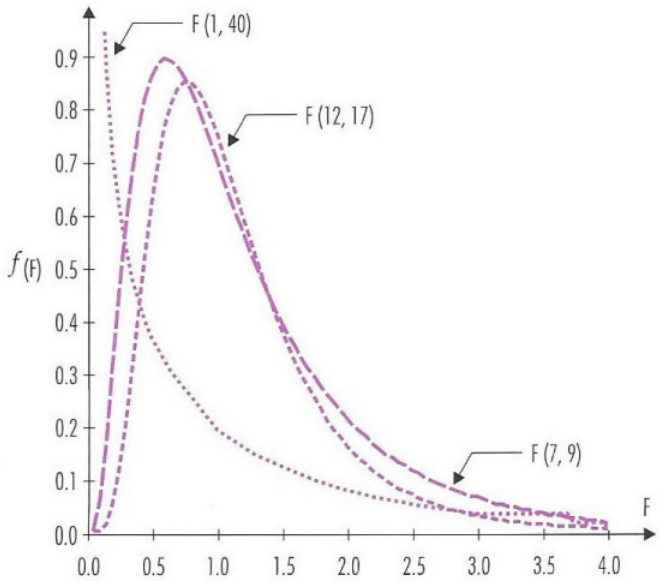
So, how do we test the hypothesis that two samples were drawn from the same population variance? We can't do a t-test, because the sample standard deviation is not normally distributed. Instead we do an F-test.

F-test steps

1. We want to know if the two population variances are equal. Null hypothesis is that $\sigma_1^2 = \sigma_2^2$.
2. Calculate the sample variance in the two samples, s_1^2 and s_2^2.
3. Form the ratio of the larger over the smaller.
4. This ratio has an F-distribution with degrees of freedom for the numerator and denominator.
5. Compare the observed F with F_{crit}, which is qf(1 - alpha/2, dfn, dfd). Note that we divide alpha by 2 for a two-tailed test.
6. Calculate the probability of the observed F as 2(1 - pf(F, dfn, dfd)). Note that we multiply by 2 if this is a two-tailed test.

F-test purpose: to compare two variances

F distribucija je teoretska distribucija vjerojatnosti, krivulja je asimetrična i različito izgleda, ovisno o n_1 i n_2 (slika 19).



You can look up the critical values of F, F_{crit} in Tablica C1 of the Vasilj textbook, page 299.

A good example of the use of the F-test is the Bumpus sparrow data. In Providence, Rhode Island, USA, in the 1890s there was a severe winter storm. Somebody collected 136 stunned sparrows and took them to Bumpus. Some of them revived, the rest died. He measured various morphological characteristics of the survivors and dead, and concluded that natural selection had eliminated individuals that were far from the average, which he considered the ideal, or the value of maximum fitness. If natural selection favors individuals closer to the mean, then the variance of the survivors should be lower than the variance of the non-survivors. We can test this hypothesis with an F-test.

Bumpus data

In females, the variance of the humerus length of survivors was 0.176, and the variance for the non-survivors was 0.434. The number of survivors was 21, and the number of non-survivors was 28. Question: is the population variance of survivors different from the population variance of non-survivors?

Exam scores, men and women

In our first midterm exam, the sample variance for women was 59.2, and the sample variance for men was 68.9. There were 14 women and 13 men taking the exam. Is the true population variance for women different from that for men?

Confidence interval for the population variance

$$\chi^2 = \frac{vs^2}{\sigma^2},$$

v is the degrees of freedom; so:

$$P(\chi^2_{0.025} < \frac{vs^2}{\sigma^2} < \chi^2_{0.975}) = 0.95$$

$$P(\frac{vs^2}{\chi^2_{0.975}} < \sigma^2 < \frac{vs^2}{\chi^2_{0.025}}) = 0.95$$

High end of 95% confidence interval: $\frac{vs^2}{\chi^2_{0.025}}$

Low end of 95% confidence interval: $\frac{vs^2}{\chi^2_{0.975}}$

Note that this method assumes that the variable itself is normally distributed, and it is sensitive to small departures from normality. So you should use it with care!

15 Analysis of variance: ANOVA

Lecture 17: Analysis of variance: ANOVA

Announcements:

- F tests and F distribution
- Analysis of variance (ANOVA)

F-test purpose: to compare two variances

F-test purpose: to compare two variances

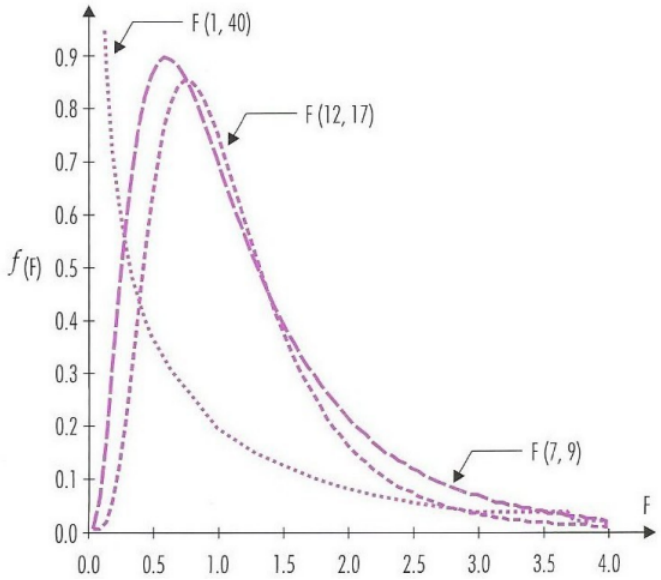

Hermon Bumpus (1862-1943)

F-test purpose: to compare two variances

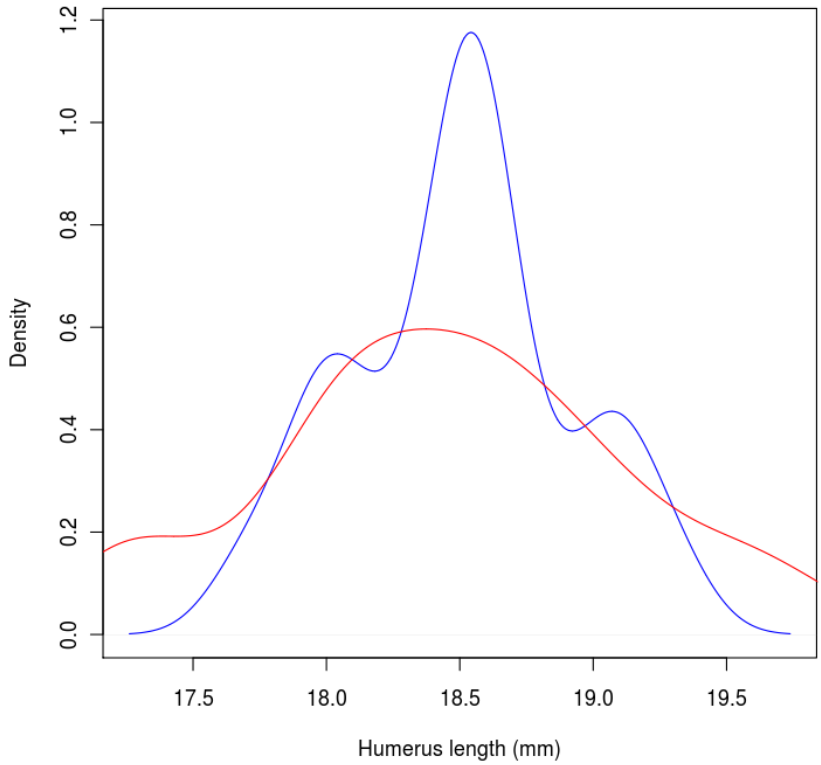
English sparrow (*Passer domesticus*)

F-test purpose: to compare two variances

F distribucija je teoretska distribucija vjerojatnosti, krivulja je asimetrična i različito izgleda, ovisno o n_1 i n_2 (slika 19).



F-test purpose: to compare two variances

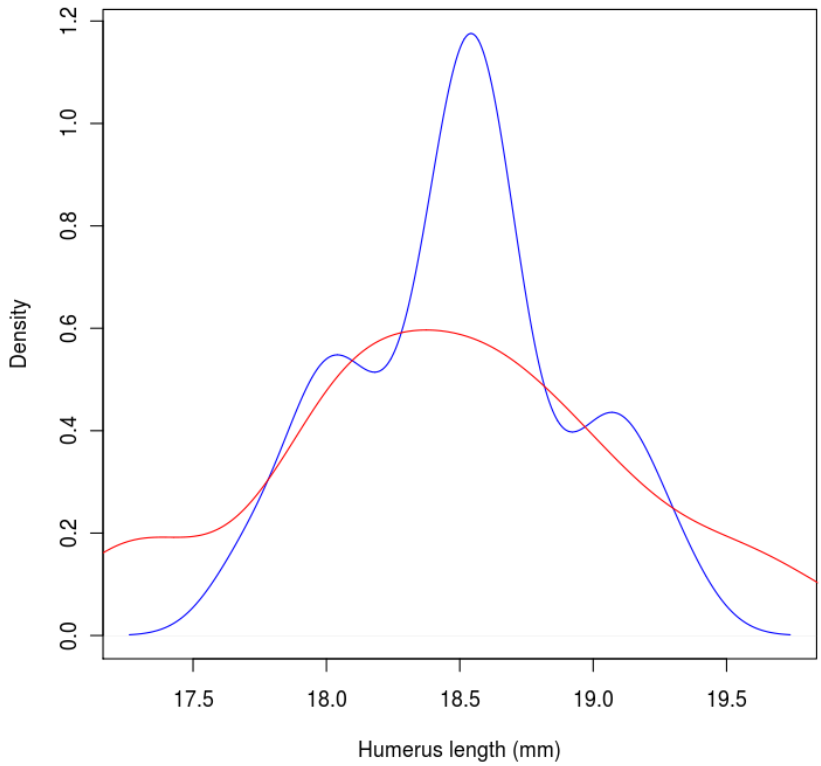


The example we gave last time was the bumpus sparrow data, where the humerus length in survivors had a significantly lower variance than the humerus length in non-survivors. The F value was 2.47, with a p = 0.04.

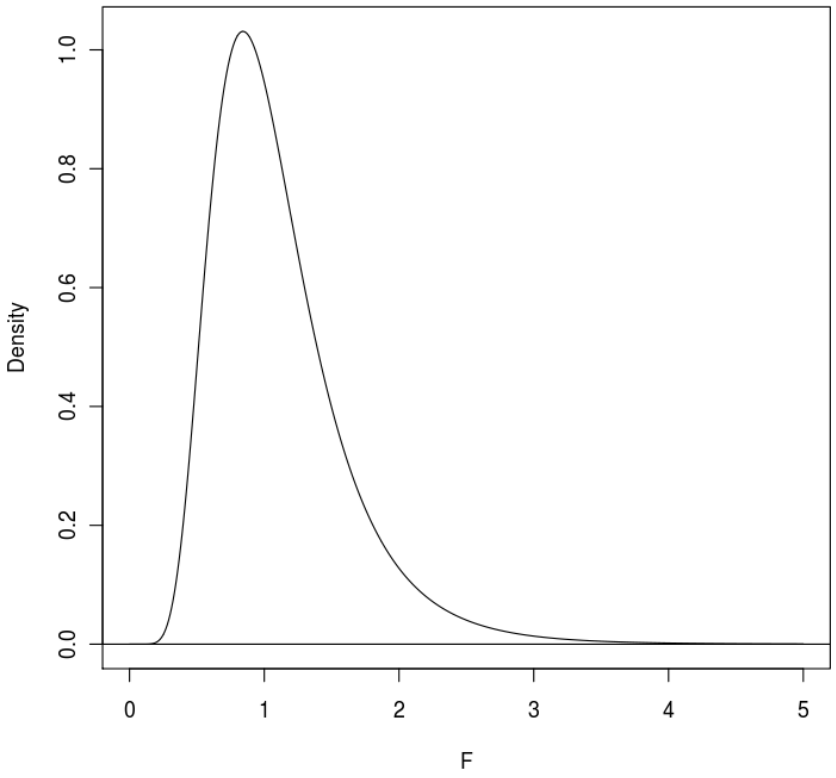
F-test steps

1. We want to know if the two population variances are equal. Null hypothesis is that $\sigma_1^2 = \sigma_2^2$.
2. Calculate the sample variance in the two samples, s_1^2 and s_2^2.
3. Form the ratio of the larger over the smaller.
4. This ratio has an F-distribution with degrees of freedom for the numerator and denominator.
5. Compare the observed F with F_{crit}, which is qf(1 - alpha/2, dfn, dfd). Note that we divide alpha by 2 for a two-tailed test.
6. Calculate the probability of the observed F as 2(1 - pf(F, dfn, dfd)). Note that we multiply by 2 if this is a two-tailed test.

F-test purpose: to compare two variances



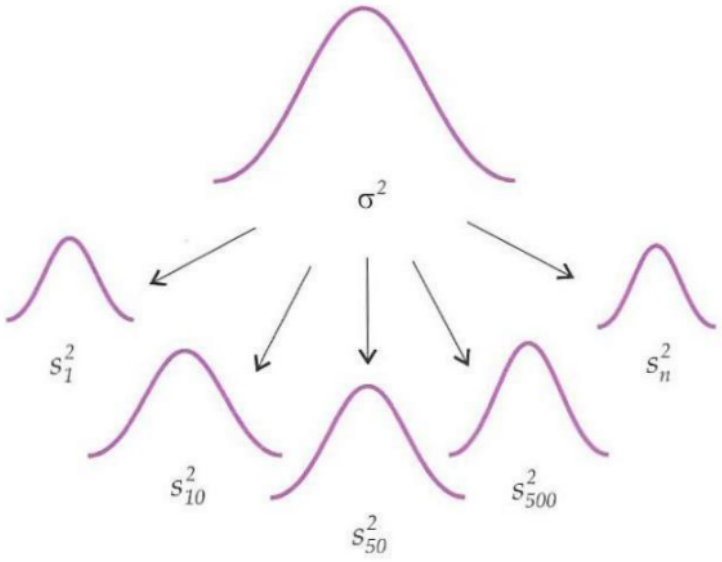
F-test purpose: to compare two variances



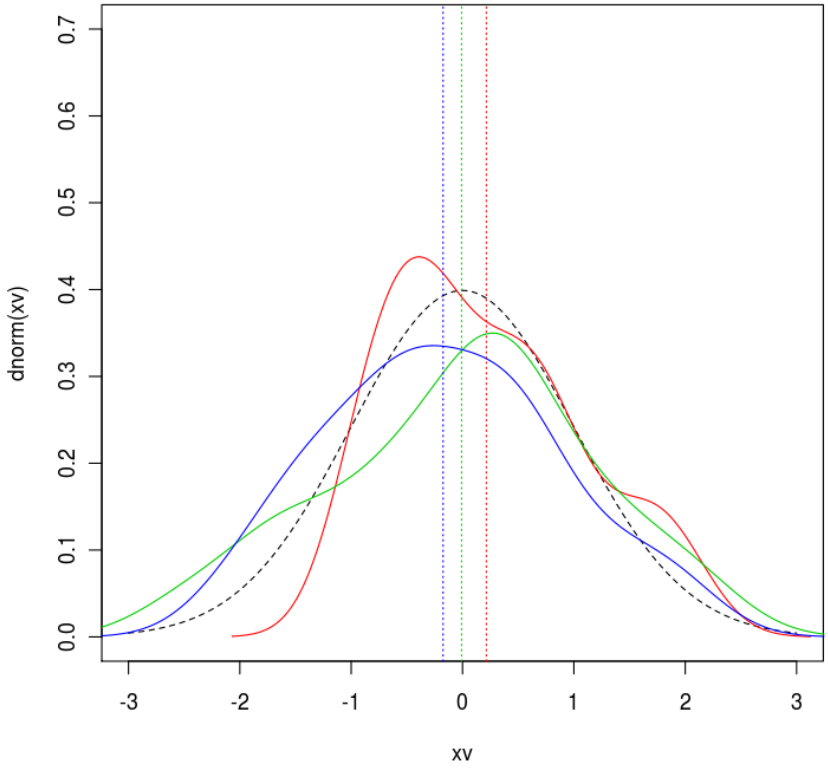
ANOVA problem

We want to test for the equality of means in more than two samples.
If we have two samples: t-test or Z-test.
If we have more than two samples: ANOVA.
In fact, ANOVA for two samples is mathematically the same as a t-test.

ANOVA problem



ANOVA problem



Single-factor ANOVA, g samples (groups), same n in all groups

1. We want to know if all population means are equal to each other. The null hypothesis is $\mu_1 = \mu_2 = \mu_2 = \ldots = \mu_k$.
2. Calculate the **observed** variance among the group means. $\text{Var}(\overline{x}_{obs}) = \frac{\sum(\overline{x}_k - \overline{\overline{x}})^2}{g-1}$.
3. Calculate the **expected** variance among group means, $\text{Var}(\overline{x}_{exp}) = \frac{s^2}{n}$.
4. Form the ratio of observed to expected:

$$F = \frac{\text{Var}(\overline{x}_{obs})}{\text{Var}(\overline{x}_{exp})} = \frac{\frac{\sum(\overline{x}_k - \overline{\overline{x}})^2}{g-1}}{\frac{s^2}{n}}.$$

Single-factor ANOVA, g samples (groups), same n in all groups

1. Compare the observed F with F_{crit}, which is qf(1 - alpha, dfn, dfd). Note that in an ANOVA this is always a one-tailed test. Numerator degrees of freedom is $g-1$. Denominator degrees of freedom is $\sum_k n_k - g$.
2. Calculate the probability of the observed F as 1 - pf(F, dfn, dfd). Note that we do not multiply by 2, we want only the right tail probability.

Single-factor ANOVA, g samples (groups), same n in all groups

1.

$$F = \frac{\text{Var}(\overline{x}_{obs})}{\text{Var}(\overline{x}_{exp})} = \frac{\frac{\sum(\overline{x}_k - \overline{\overline{x}})^2}{g-1}}{\frac{s^2}{n}} = \frac{n\sum(\overline{x}_k - \overline{\overline{x}})^2/(g-1)}{s^2}.$$

$$F = \frac{\text{SS}_g/\text{df}_g}{s^2}.$$

2.
$$s^2 = \frac{\sum(x_1 - \overline{x}_1)^2 + \ldots + \sum(x_g - \overline{x}_g)^2}{\sum n - g}.$$

3.
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{\mathrm{SS}_w/\mathrm{df}_w} = \frac{\mathrm{MS}_g}{\mathrm{MS}_w}.$$

Single-factor ANOVA, g samples (groups), different n_k in all groups

1.
$$F = \frac{\sum n_k(\overline{x}_k - \overline{\overline{x}})^2/(g-1)}{s^2}.$$
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{s^2}.$$

2.
$$s^2 = \frac{\sum(x_1 - \overline{x}_1)^2 + \ldots + \sum(x_g - \overline{x}_g)^2}{\sum n_k - g}.$$

3.
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{\mathrm{SS}_w/\mathrm{df}_w} = \frac{\mathrm{MS}_g}{\mathrm{MS}_w}.$$

The proportion variation explained is called the coefficient of determination, denoted r^2, and is equal to $SS_g/(SS_g + SS_w)$. Generally these results are presented in an ANOVA table, with columns SS, df, MS, F, and sometimes a box also for probability. The rows are groups and within, and sometimes you can add a row indicating the totals. The total sums of squares $SS_t = SS_g + SS_w$. The total degrees of freedom $df_t = df_g + df_w$.

16 Multiple comparisons and 2-factor ANOVA

Lecture 18: 2-factor ANOVA

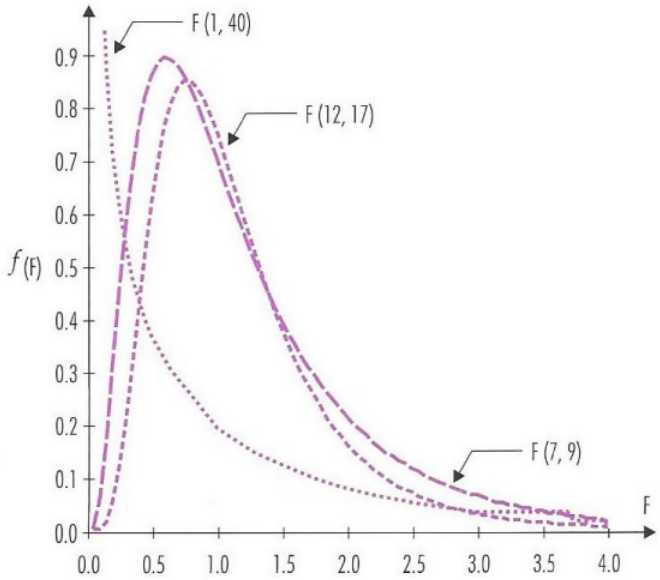
Announcements:

- Quiz 9 is up and ready, due Wednesday.
- Midterm Exam 2 will be Tuesday, December 20, 2011, in the lab.
- Midterm Exam 2 will cover quizzes 6, 7, 8, 9, and 10.
- Midterm 2 Practice Exam is up.

F-test purpose: to compare two variances

F-test purpose: to compare two variances

F distribucija je teoretska distribucija vjerojatnosti, krivulja je asimetrična i različito izgleda, ovisno o n_1 i n_2 (slika 19).



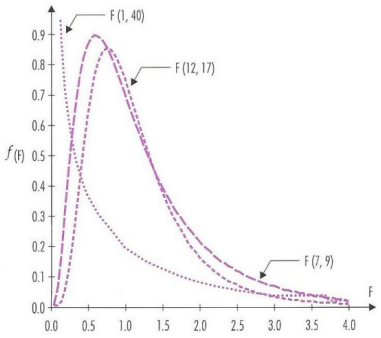
The F-test is used to test for the equality of population variances, using the ratio of sample variances, which follows an F distribution. The F-test is used in the analysis of variance (ANOVA), to test for the equality of means, not variances, by comparing the observed variance among sample means to the predicted variance among sample means if all samples were drawn from the same population.

Remember, the purpose of ANOVA is to explain variation. We have a variable, and we see that it varies from individual to individual, or from place to place, or from time to time, and we want to know what the causes of that variation are. The variable we're interested in is called the "response" variable. It is the variable that responds to some other variable, so that its variation is "explained" by variation in the other variable. The other variable, that explains the first variable is called the "predictor" variable, because it explains the variable in our variable of interest. So, for example, we see insect damage of a plant. Different plants have different degrees of insect damage. Why do they vary? Maybe it is because different plants have higher numbers of parasites than others. So in this example, the damage is the response variable, and its variation is partly influenced by the number of parasites, which is the predictor variable. So we say that the predictor explains variation in the response. In an ANOVA we can calculate exactly how much of the variation in the response variable is explained by the predictor variable.

So far our predictor variable has been categorical. A categorical variable is a variable that assigns each observation to a category. Examples are sex (categories are male or female), variety (categories A, B, C, etc.), location, site, season. A categorical predictor variable is called a "factor". Each of the category choices within the factor are called "levels" of the factor. So the factor sex for example has two levels, male and female. The factor season has four levels, winter, spring, summer, and fall. In R, you need to make sure that when you label categories with numbers, those numbers are treated as factor levels rather than numbers. For example if you label the sexes 1 and 2 you need to make sure R sees those as factor levels and not the numbers 1 and 2.

F-test purpose: to compare two variances

F distribucija je teoretska distribucija vjerojatnosti, krivulja je asimetrična i različito izgleda, ovisno o n_1 i n_2 (slika 19).



ANOVA purpose is to compare means of populations, not variances. But any difference can be expressed as a variance. So the difference among means can be tested with an F test.

The single-factor ANOVA can be thought of as a multi-sample t-test. If there are only two samples, then the ANOVA is mathematically equivalent to a two-sample t-test assuming equal variances in the two populations.

Single-factor ANOVA, g samples (groups), different n_k in all groups

1.
$$F = \frac{\sum n_k(\overline{x}_k - \overline{\overline{x}})^2/(g-1)}{s^2}.$$
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{s^2}.$$

2.
$$s^2 = \frac{\sum(x_1 - \overline{x}_1)^2 + \ldots + \sum(x_g - \overline{x}_g)^2}{\sum n_k - g}.$$

3.
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{\mathrm{SS}_w/\mathrm{df}_w} = \frac{\mathrm{MS}_g}{\mathrm{MS}_w}.$$

4. Note that $s^2 = \mathrm{MS}_w$. This is the "pooled variance" within groups.

Calculating pooled variance s^2

The pooled variance is the weighted mean variance of the groups, where the weights are the degrees of freedom.

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 + \ldots + (n_g-1)s_g^2}{\sum n_k - g}$$
$$\mathrm{MS}_w = \text{pooled variance} = s^2 = \frac{SS_w}{df_w}$$
$$SS_w = \sum(x_1 - \overline{x}_1)^2 + \ldots + \sum(x_g - \overline{x}_g)^2$$
$$df_w = \sum n_k - g$$

In the single-factor ANOVA, we have just one response variable and one predictor variable. The response variable is a continuous variable (like height or weight), and the predictor variable is a factor (like sex or species or treatment).

In an ANOVA, the result is an analysis of variance table, which always has the same components.

Single-factor ANOVA

```
Analysis of Variance Table

Response: Prevalence
          Df Sum Sq Mean Sq F value    Pr(>F)
Alcohol    3 3.4510 1.15032  12.617 6.914e-07 ***
Residuals 84 7.6584 0.09117
```

Single-factor ANOVA

1. $$F = \frac{\mathrm{MS}_g}{\mathrm{MS}_w}.$$
2. $$\mathrm{MS}_g = \frac{\mathrm{SS}_g}{\mathrm{df}_g}$$
3. $$\mathrm{SS}_g = \sum n_k(\overline{x}_k - \overline{\overline{x}})^2$$
4. $$\mathrm{df}_g = g - 1$$
5. $$\mathrm{MS}_w = \text{pooled variance} = s^2 = \frac{SS_w}{df_w}$$

For example, let's take the dataset data(chickwts) in R. This is the weight of chicks after some length of time fed on different diets. We want to know if there is any difference among the diets in the weight of chicks. We answer this question with a single-factor ANOVA: anova(lm(weight ~ feed)). We see in the ANOVA that there is a significant difference. The next question is: which diets are actually different from each other? We can visualize the differences among diets with the command boxplot(weight ~ feed), but the plot doesn't answer the question. We can answer this question by doing all possible t-tests for all possible pairs of diets. But for each t-test, the false positive rate is 0.05. If we do many t-tests, the probability that at least one of the tests gives a false positive is greater than 0.05. So doing a large number of tests, looking for a significant result, is a "fishing expedition" that greatly increases the false positive rate. We can correct this problem by adjusting the probability of the test upwards based on the number of tests performed, and still using the 0.05 criterion for significance. This adjusted probability is used in the Tukey Honestly Significant Difference method for performing multiple t-tests.

Multiple comparisons problem

- Multiple t-tests inflate the false positive rate!
- Therefore, we need to adjust for the number of comparisons.
- The Tukey Honestly Significant Difference tests makes this adjustment.
- In R, the function is TukeyHSD().
- First do a = aov(response ~ predictor), then TukeyHSD(a, "predictor").
- Or, just TukeyHSD(aov(response ~ predictor)) to do all possible comparisons within each predictor.

For the chick weight data, we do:

```
data(chickwts)
attach(chickwts)
a = aov(weight \string~ feed)
Tukey(a, ‘‘feed’’)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight \string~ feed)

$feed
                          diff         lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

This shows us the difference in mean weight between each feed type, the 95% confidence interval for the difference (adjusted for the multiple comparisons) and the probability in the t-test (again, adjusted for the multiple comparisons). Here we find out that the largest significant difference in weight is between sunflower and horsebean diets. Some diets are not significantly different from each other, as for example sunflower and casein.

2-factor ANOVA: calculations

We do three hypothesis tests in 2-factor ANOVA. We test for:

- the main effect of factor 1.
- the main effect of factor 2.
- the interaction between the two factors.

Take, for example, two factors, very commonly used: sex, and drugs. Often when you take a medication, there are side effects. For example, if you take steroids. The side effects are different in men and women. For example, women taking steroids often develop a deeper voice.

2-factor ANOVA: calculations

- The mean squares for each effect are the same as for the single-factor ANOVA.
- The within mean squares is different from the single-factor ANOVA: it is the pooled variance within each cell, rather than within each effect.
- The interaction is calculated as a difference.

2-factor ANOVA: calculations

- The mean squares for each effect are the same as for the single-factor ANOVA.
- The within mean squares is different from the single-factor ANOVA: it is the pooled variance within each cell, rather than within each effect.
- The interaction is calculated as a difference.

As an example, let's do a 2-factor ANOVA for voice deepness, with factors steroid use and sex.

Here are the data: no steroids, female: mean is 3, variance is 3, sample size is 20. yes steroids, female: mean is 6, variance is 3, sample size is 20. no steroids, male: mean is 7, variance is 2, sample size is 20. yes steroids, male: mean is 8, variance is 2, sample size is 20.

How do we calculate the ANOVA table for these data, assuming we're testing for a drug effect, a sex effect, and the interaction between the two? See the tablet file for the calculations.

Some additional calculations are generally interesting: the coefficient of determination, which is the proportion of the total variation that is explained by the effect. This is just the sum of squares for the effect divided by the total sum of squares. The total sum of squares is the sum of all the sums of squares (factor1, factor2, interaction, within). Also, when we have a 2-by-2 table, we can actually calculate the interaction. This is effect of factor 1 within the first level of factor two, minus the effect of factor 1 within the second level of factor 2. In the steroid example the sample or observed interaction is (3-6) - (7-8) = (-3) - (-1) = -2.

What are the reasons for doing a 2-factor ANOVA, rather than two single-factor ANOVAs? First, if there is an interaction, then you explain more of the variation. Second, if there is an interaction, then your interpretation of the main effect is different, it becomes conditional. And third, whether there is an interaction or not, then the mean squares within is smaller if you do a 2-factor ANOVA, which results in a higher power of the test.

2-factor ANOVA: coefficient of determination, r^2

```
Analysis of Variance Table

Response: Prevalence
           Df Sum Sq Mean Sq F value    Pr(>F)
Alcohol     3 3.4510 1.15032  12.617 6.914e-07 ***
Residuals 84 7.6584 0.09117
```

In the above ANOVA table, the coefficient of determination is 3.45/(3.45 + 7.66) = 0.31. In other words, 31% of the variation in the response variable is explained by the predictor. Variation is synonymous with "sum of squares". To calculate the coefficient of determination (r^2) for a factor, we divide its sum of squares with the total sum of squares. The total sum of squares is the factor sum of squares plus the within sum of squares.

2-factor ANOVA: rationale

Why do a 2-factor ANOVA rather than two 1-factor ANOVAs?

- If an interaction exists, then the interpretation of the main effect is different.
- If an interaction exists, then we can demonstrate it and increase the proportion of variation that we have explained (i.e. increase the coefficient of determination, r^2).
- If there is a large effect of both of the factors, then including both in the same analysis greatly reduces the within mean squares, which increases the power of the test for demonstrating both main effects.

Lecture 19: 2-factor ANOVA

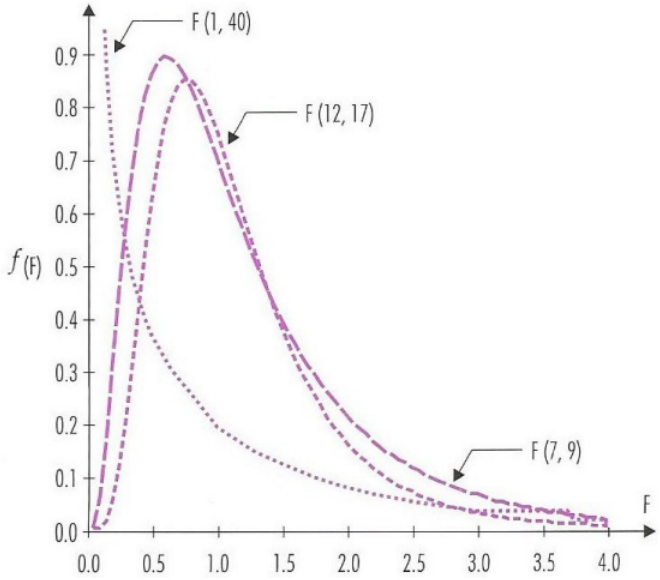
Announcements:

- Quiz 9 is up and ready, due Wednesday.
- Midterm Exam 2 will be Tuesday, December 20, 2011, in the lab.
- Midterm Exam 2 will cover quizzes 6, 7, 8, 9, and 10.
- Midterm 2 Practice Exam will be up soon.

F-test purpose: to compare two variances

F-test purpose: to compare two variances

F distribucija je teoretska distribucija vjerojatnosti, krivulja je asimetrična i različito izgleda, ovisno o n_1 i n_2 (slika 19).



The F-test is used to test for the equality of population variances, using the ratio of sample variances, which follows an F distribution. The F-test is used in the analysis of variance (ANOVA), to test for the equality of means, not variances, by comparing the observed variance among sample means to the predicted variance among sample means if all samples were drawn from the same population.

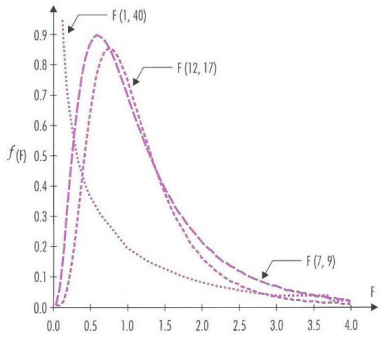
Remember, the purpose of ANOVA is to explain variation. We have a variable, and we see that it varies from individual to individual, or from place to place, or from time to time, and we want to know what the causes of that variation are. The variable we're interested in is called the "response" variable. It is the variable that responds to some other variable, so that its variation is "explained" by variation in the other variable. The other variable, that explains the first variable is called the "predictor" variable, because it explains the variable in our variable of interest. So, for example, we see insect damage of a plant. Different plants have different degrees of insect damage. Why do they vary? Maybe it is because different plants have higher numbers of parasites than others. So in this example, the damage is the response variable, and its variation is partly influenced by the number of parasites, which is the predictor variable. So we say that the predictor explains variation in the response. In an ANOVA we can calculate exactly how much of the variation in the response variable is explained by the predictor variable.

So far our predictor variable has been categorical. A categorical variable is a variable that assigns each observation to a category. Examples are sex (categories are male or female), variety (categories A, B, C, etc.), location, site, season. A categorical predictor variable is called a "factor". Each of the category choices within the factor are called "levels" of the factor. So the factor sex for example has two levels, male and female. The factor season has four levels, winter, spring, summer, and fall. In R, you need to make sure that when you label categories with numbers, those numbers are treated as factor

levels rather than numbers. For example if you label the sexes 1 and 2 you need to make sure R sees those as factor levels and not the numbers 1 and 2.

F-test purpose: to compare two variances

F distribucija je teoretska distribucija vjerojatnosti, krivulja je asimetrična i različito izgleda, ovisno o n_1 i n_2 (slika 19).



ANOVA purpose is to compare means of populations, not variances. But any difference can be expressed as a variance. So the difference among means can be tested with an F test.

The single-factor ANOVA can be thought of as a multi-sample t-test. If there are only two samples, then the ANOVA is mathematically equivalent to a two-sample t-test assuming equal variances in the two populations.

Single-factor ANOVA, g samples (groups), different n_k in all groups

1.
$$F = \frac{\sum n_k(\overline{x}_k - \overline{\overline{x}})^2/(g-1)}{s^2}.$$
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{s^2}.$$

2.
$$s^2 = \frac{\sum(x_1 - \overline{x}_1)^2 + \ldots + \sum(x_g - \overline{x}_g)^2}{\sum n_k - g}.$$

3.
$$F = \frac{\mathrm{SS}_g/\mathrm{df}_g}{\mathrm{SS}_w/\mathrm{df}_w} = \frac{\mathrm{MS}_g}{\mathrm{MS}_w}.$$

4. Note that $s^2 = \mathrm{MS}_w$. This is the "pooled variance" within groups.

Calculating pooled variance s^2

The pooled variance is the weighted mean variance of the groups, where the weights are the degrees of freedom.

$$s^2 = \frac{(n_1-1)s_1^2 + (n_2-1)s_2^2 + (n_3-1)s_3^2 + \ldots + (n_g-1)s_g^2}{\sum n_k - g}$$
$$\mathrm{MS}_w = \text{pooled variance} = s^2 = \frac{SS_w}{df_w}$$
$$SS_w = \sum(x_1 - \overline{x}_1)^2 + \ldots + \sum(x_g - \overline{x}_g)^2$$
$$df_w = \sum n_k - g$$

In the single-factor ANOVA, we have just one response variable and one predictor variable. The response variable is a continuous variable (like height or weight), and the predictor variable is a factor (like sex or species or treatment).

In an ANOVA, the result is an analysis of variance table, which always has the same components.

Single-factor ANOVA

```
Analysis of Variance Table

Response: Prevalence
          Df Sum Sq Mean Sq F value    Pr(>F)
Alcohol    3 3.4510 1.15032  12.617 6.914e-07 ***
Residuals 84 7.6584 0.09117
```

Single-factor ANOVA

1. $$F = \frac{\mathrm{MS}_g}{\mathrm{MS}_w}.$$
2. $$\mathrm{MS}_g = \frac{\mathrm{SS}_g}{\mathrm{df}_g}$$
3. $$\mathrm{SS}_g = \sum n_k(\overline{x}_k - \overline{\overline{x}})^2$$
4. $$\mathrm{df}_g = g - 1$$
5. $$\mathrm{MS}_w = \text{pooled variance} = s^2 = \frac{SS_w}{df_w}$$

For example, let's take the dataset data(chickwts) in R. This is the weight of chicks after some length of time fed on different diets. We want to know if there is any difference among the diets in the weight of chicks. We answer this question with a single-factor ANOVA: anova(lm(weight ~ feed)). We see in the ANOVA that there is a significant difference. The next question is: which diets are actually different from each other? We can visualize the differences among diets with the command boxplot(weight ~ feed), but the plot doesn't answer the question. We can answer this question by doing all possible t-tests for all possible pairs of diets. But for each t-test, the false positive rate is 0.05. If we do many t-tests, the probability that at least one of the tests gives a false positive is greater than 0.05. So doing a large number of tests, looking for a significant result, is a "fishing expedition" that greatly increases the false positive rate. We can correct this problem by adjusting the probability of the test upwards based on the number of tests performed, and still using the 0.05 criterion for significance. This adjusted probability is used in the Tukey Honestly Significant Difference method for performing multiple t-tests.

Multiple comparisons problem

- Multiple t-tests inflate the false positive rate!
- Therefore, we need to adjust for the number of comparisons.
- The Tukey Honestly Significant Difference tests makes this adjustment.
- In R, the function is TukeyHSD().
- First do a = aov(response ~ predictor), then TukeyHSD(a, "predictor").
- Or, just TukeyHSD(aov(response ~ predictor)) to do all possible comparisons within each predictor.

For the chick weight data, we do:

```
data(chickwts)
attach(chickwts)
a = aov(weight \string~ feed)
Tukey(a, ‘‘feed’’)

  Tukey multiple comparisons of means
    95% family-wise confidence level

Fit: aov(formula = weight \string~ feed)

$feed
                          diff         lwr       upr     p adj
horsebean-casein    -163.383333 -232.346876 -94.41979 0.0000000
linseed-casein      -104.833333 -170.587491 -39.07918 0.0002100
meatmeal-casein      -46.674242 -113.906207  20.55772 0.3324584
soybean-casein       -77.154762 -140.517054 -13.79247 0.0083653
sunflower-casein       5.333333  -60.420825  71.08749 0.9998902
linseed-horsebean     58.550000  -10.413543 127.51354 0.1413329
meatmeal-horsebean   116.709091   46.335105 187.08308 0.0001062
soybean-horsebean     86.228571   19.541684 152.91546 0.0042167
sunflower-horsebean  168.716667   99.753124 237.68021 0.0000000
meatmeal-linseed      58.159091   -9.072873 125.39106 0.1276965
soybean-linseed       27.678571  -35.683721  91.04086 0.7932853
sunflower-linseed    110.166667   44.412509 175.92082 0.0000884
soybean-meatmeal     -30.480519  -95.375109  34.41407 0.7391356
sunflower-meatmeal    52.007576  -15.224388 119.23954 0.2206962
sunflower-soybean     82.488095   19.125803 145.85039 0.0038845
```

This shows us the difference in mean weight between each feed type, the 95% confidence interval for the difference (adjusted for the multiple comparisons) and the probability in the t-test (again, adjusted for the multiple comparisons). Here we find out that the largest significant difference in weight is between sunflower and horsebean diets. Some diets are not significantly different from each other, as for example sunflower and casein.

2-factor ANOVA: calculations

We do three hypothesis tests in 2-factor ANOVA. We test for:

- the main effect of factor 1.
- the main effect of factor 2.
- the interaction between the two factors.

Take, for example, two factors, very commonly used: sex, and drugs. Often when you take a medication, there are side effects. For example, if you take steroids. The side effects are different in men and women. For example, women taking steroids often develop a deeper voice.

2-factor ANOVA: calculations

- The mean squares for each effect are the same as for the single-factor ANOVA.
- The within mean squares is different from the single-factor ANOVA: it is the pooled variance within each cell, rather than within each effect.
- The interaction is calculated as a difference.

2-factor ANOVA: calculations

- The mean squares for each effect are the same as for the single-factor ANOVA.
- The within mean squares is different from the single-factor ANOVA: it is the pooled variance within each cell, rather than within each effect.
- The interaction is calculated as a difference.

As an example, let's do a 2-factor ANOVA for voice deepness, with factors steroid use and sex.

Here are the data: no steroids, female: mean is 3, variance is 3, sample size is 20. yes steroids, female: mean is 6, variance is 3, sample size is 20. no steroids, male: mean is 7, variance is 2, sample size is 20. yes steroids, male: mean is 8, variance is 2, sample size is 20.

How do we calculate the ANOVA table for these data, assuming we're testing for a drug effect, a sex effect, and the interaction between the two? See the tablet file for the calculations.

Some additional calculations are generally interesting: the coefficient of determination, which is the proportion of the total variation that is explained by the effect. This is just the sum of squares for the effect divided by the total sum of squares. The total sum of squares is the sum of all the sums of squares (factor1, factor2, interaction, within). Also, when we have a 2-by-2 table, we can actually calculate the interaction. This is effect of factor 1 within the first level of factor two, minus the effect of factor 1 within the second level of factor 2. In the steroid example the sample or observed interaction is (3-6) - (7-8) = (-3) - (-1) = -2.

What are the reasons for doing a 2-factor ANOVA, rather than two single-factor ANOVAs? First, if there is an interaction, then you explain more of the variation. Second, if there is an interaction, then your interpretation of the main effect is different, it becomes conditional. And third, whether there is an interaction or not, then the mean squares within is smaller if you do a 2-factor ANOVA, which results in a higher power of the test.

2-factor ANOVA: coefficient of determination, r^2

```
Analysis of Variance Table

Response: Prevalence
           Df Sum Sq Mean Sq F value    Pr(>F)
Alcohol     3 3.4510 1.15032  12.617 6.914e-07 ***
Residuals 84 7.6584 0.09117
```

In the above ANOVA table, the coefficient of determination is 3.45/(3.45 + 7.66) = 0.31. In other words, 31% of the variation in the response variable is explained by the predictor. Variation is synonymous with "sum of squares". To calculate the coefficient of determination (r^2) for a factor, we divide its sum of squares with the total sum of squares. The total sum of squares is the factor sum of squares plus the within sum of squares.

2-factor ANOVA: rationale

Why do a 2-factor ANOVA rather than two 1-factor ANOVAs?

- If an interaction exists, then the interpretation of the main effect is different.
- If an interaction exists, then we can demonstrate it and increase the proportion of variation that we have explained (i.e. increase the coefficient of determination, r^2).
- If there is a large effect of both of the factors, then including both in the same analysis greatly reduces the within mean squares, which increases the power of the test for demonstrating both main effects.

Lecture 20 was just making jpg files of the quiz 9 questions and showing the class how to do each of the questions.

17 Covariance, correlation, and linear regression

Lecture 21: Covariance, correlation, and linear regression

Announcements:

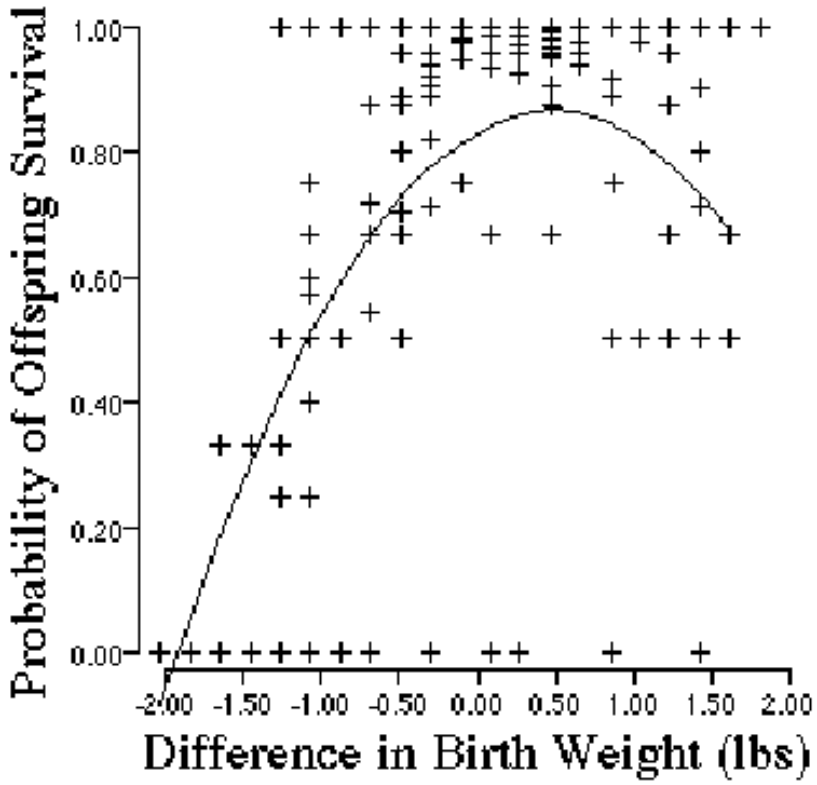
- Reading: Chapter 7 in Vasilj.
- Next on the menu: Covariance, correlation, and linear regression

Problem. Two continuous variables: are they related?

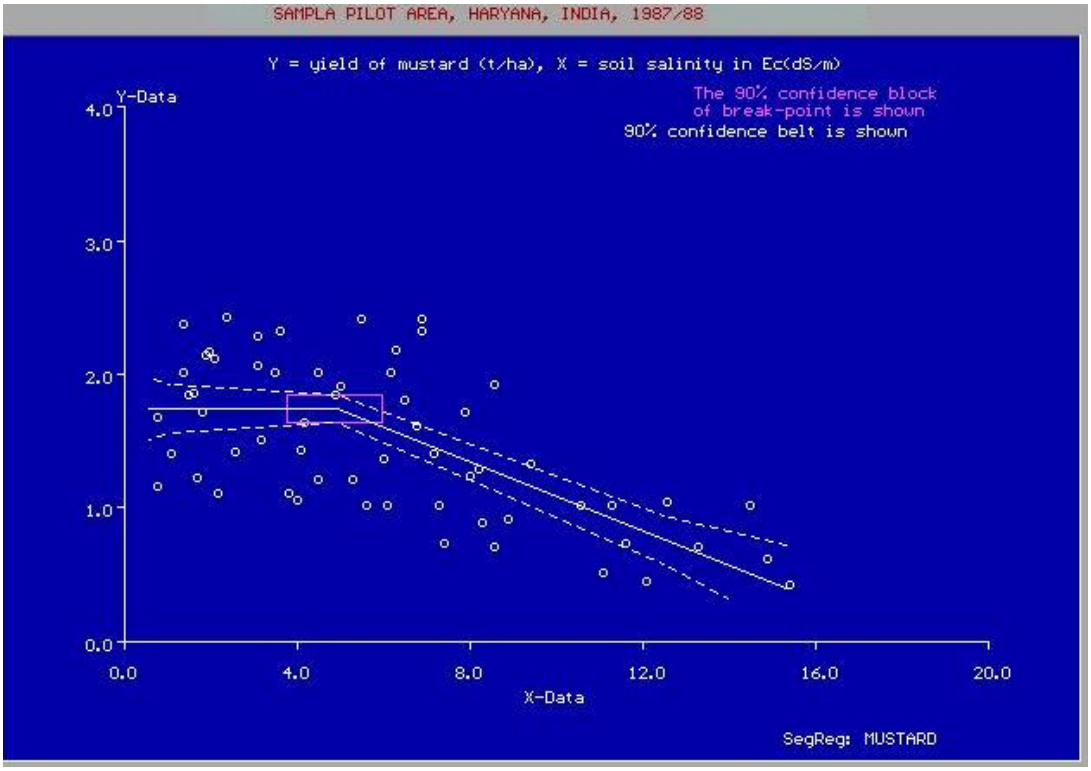
Examples

- Parents versus offspring.
- Growth curves (organs, organisms, populations).
- Allometric relationships.
- Physiological relationships.
- Ecological relationships.

Regression of traits on fitness



Ecological regressions can be messy



Correlation versus causation

Storks 

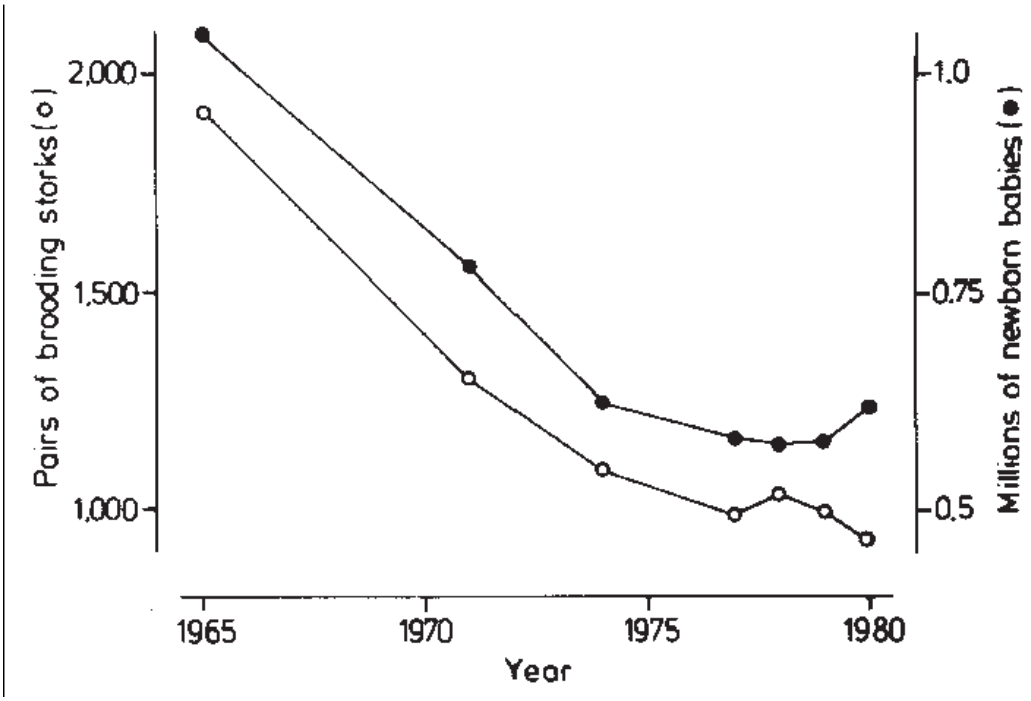
Correlation versus causation

Babies

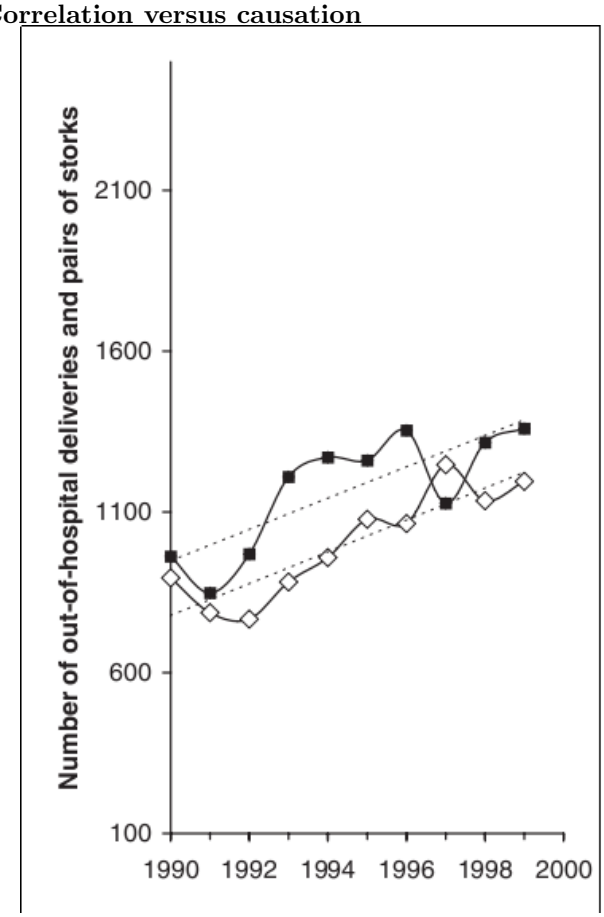
Correlation versus causation

Storks and babies?

Correlation versus causation



Correlation versus causation



Covariance

1. $Cov(x, y) = E[(x - \overline{x})(y - \overline{y})]$

2.

$$Cov(x, y) = \frac{\sum_i [(x_i - \overline{x})(y_i - \overline{y})]}{n - 1}$$

Properties of the covariance

1. $\text{Cov}(x, a) = 0$, where a is any constant.
2. $\text{Cov}(x, x) = \text{Var}(x)$, for any variable x.
3. $\text{Cov}(x, y) = \text{Cov}(y, x)$, for any variables x and y.
4. $\text{Cov}(x, bx) = b\text{Var}(x)$, for any variables x and y and constant b.
5. $\text{Cov}(ax, by) = ab\text{Cov}(x, y)$, for any constants a and b.
6. $\text{Cov}(x + a, y + b) = \text{Cov}(x, y)$, for any constants a and b.
7. If x and y are independent, then $\text{Cov}(x, y) = 0$.

Calculating the slope and intercept of the best-fit line

The best fit line is the line that minimizes the residual sum of squares.

1.

$$\text{Slope} = b = \frac{\text{Cov}(x, y)}{\text{Var}(x)}.$$

2. Intercept $= a = \overline{y} - b\overline{x}$.

Correlation versus covariance

Correlation coefficient is the scaled covariance.

1. The maximum covariance between x and y is $s_x s_y$.
2.

$$\text{Cor}(x, y) = r = \frac{\text{Cov}(x, y)}{s_x s_y}.$$

3. The correlation coefficient lies between -1 and +1.

Strength of the correlation coefficient

korelacijski koeficijent (r)	jačina korelacije
0.00 – 0.10	nema
0.10 – 0.25	vrlo slaba
0.25 – 0.40	slaba
0.40 – 0.50	srednja
0.50 – 0.75	jaka
0.75 – 0.90	vrlo jaka
0.90 – 1.00	potpuna

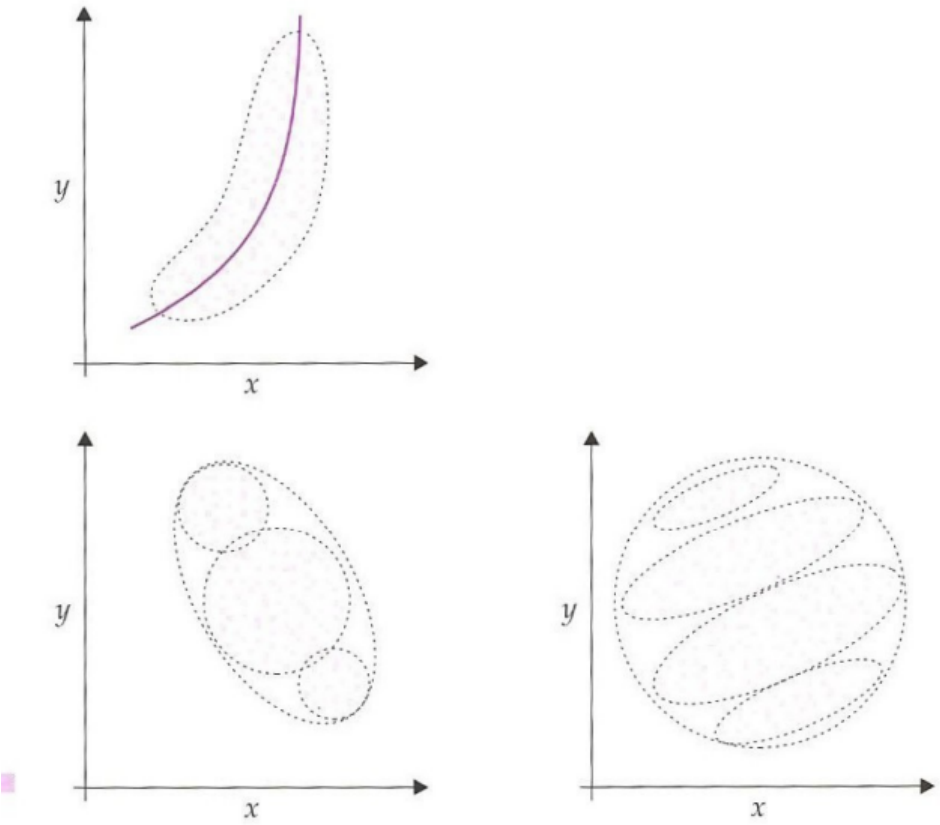
Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.
3. Calculate the line sum of squares $SS_x = \sum(\hat{y} - \overline{y})^2$.
4. Calculate the corresponding residual and line mean squares: $MS_{res} = SS_{res}/df_{res}$; and $MS_x = SS_x/df_x$ $df_{res} = n - 2; df_x = 1$
5. Form the F ratio and calculate its tail probability $F = MS_{res}/MS_x$.
6. Or let R do it all with anova(lm(y ~ x)).

Assumptions of linear regression

1. The x values are measured without any error
2. The relationship between x and y is linear
3. The residuals are normally distributed
4. The variance is constant (regardless of the value of x)

Interpretation of correlation



Lecture 23: Regression

Announcements:

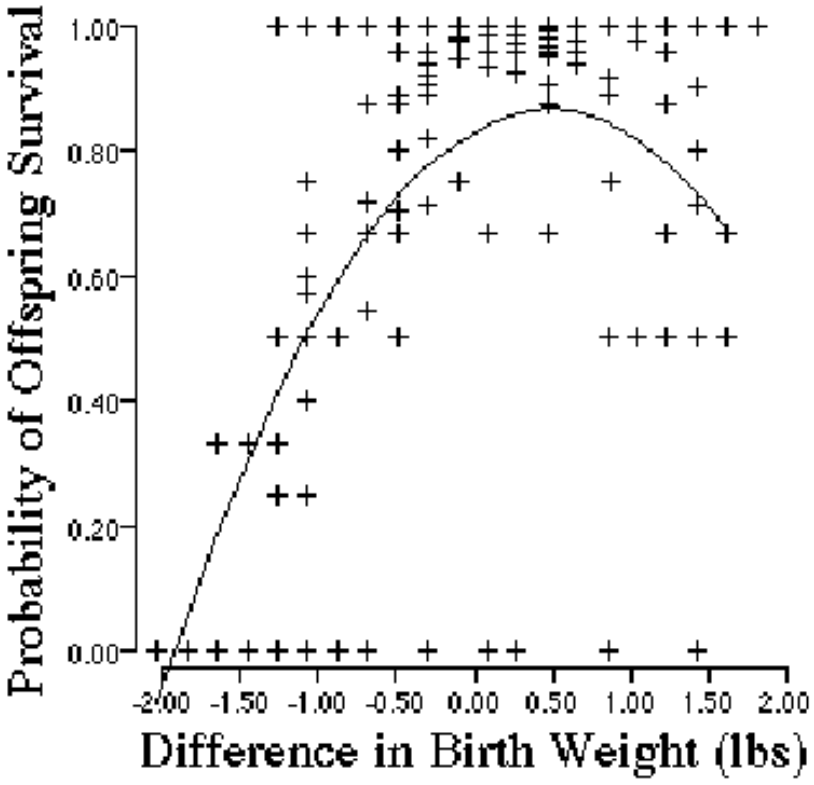
- Reading: Chapter 7 in Vasilj.
- Third colloquium exam on Tuesday, January 24, 14:00 to 18:00.
- Today: Regression review, nonlinear regression

Problem. Two continuous variables: are they related?

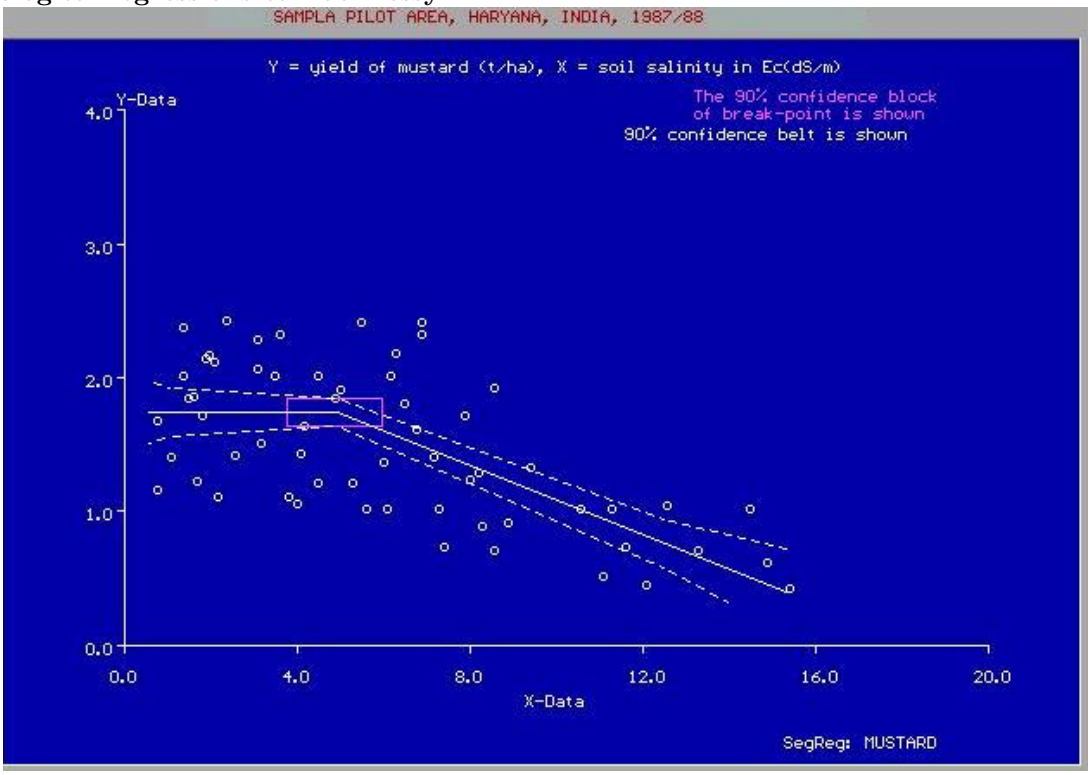
Examples

- Parents versus offspring.
- Growth curves (organs, organisms, populations).
- Allometric relationships.
- Regression of traits on fitness.
- Physiological relationships.
- Ecological relationships.

Regression of traits on fitness



Ecological regressions can be messy



Covariance

1. $Cov(x, y) = E[(x - \overline{x})(y - \overline{y})]$

2.

$$Cov(x, y) = \frac{\sum_i [(x_i - \overline{x})(y_i - \overline{y})]}{n - 1}$$

Properties of the covariance

1. $\text{Cov}(x, a) = 0$, where a is any constant.
2. $\text{Cov}(x, x) = \text{Var}(x)$, for any variable x.
3. $\text{Cov}(x, y) = \text{Cov}(y, x)$, for any variables x and y.
4. $\text{Cov}(x, bx) = b\text{Var}(x)$, for any variables x and y and constant b.
5. $\text{Cov}(ax, by) = ab\text{Cov}(x, y)$, for any constants a and b.
6. $\text{Cov}(x + a, y + b) = \text{Cov}(x, y)$, for any constants a and b.
7. If x and y are independent, then $\text{Cov}(x, y) = 0$.

Simple linear regression

1. Calculate the best-fit straight line $\hat{y} = a + bx$.
2. Calculate the residual sum of squares. $SS_{res} = \sum(y - \hat{y})^2$.
3. Calculate the line sum of squares $SS_x = \sum(\hat{y} - \overline{y})^2$.
4. Calculate the corresponding residual and line mean squares: $MS_{res} = SS_{res}/df_{res}$; and $MS_x = SS_x/df_x$ $df_{res} = n - 2; df_x = 1$
5. Form the F ratio and calculate its tail probability $F = MS_{res}/MS_x$.
6. Or let R do it all with anova(lm(y ~ x)).

Simple linear regression: shortcut

Let's say you don't have the raw data, just the covariances, means, variances, and sample size. Here's how you calculate the ANOVA table:

1. $df_t = n - 1$, $df_{res} = n - 2$, $df_x = 1$.
2. Since $var(y) = MS_t = SS_t/df_t = SS_t/(n - 1)$, then $SS_t = var(y)(n - 1)$.
3. Calculate r, the correlation coefficient: $r = cov(x, y)/(s_x s_y)$.
4. Since $r^2 = SS_x/SS_t$, we get $SS_x = r^2 SS_t$.
5. Since $1 - r^2 = SS_{res}/SS_t$, $SS_{res} = (1 - r^2)SS_t$.
6. Now the mean squares are easy: $MS_x = SS_x/1$, $MS_{res} = SS_{res}/df_{res}$.
7. Form the F ratio and calculate its tail probability $F = MS_{res}/MS_x$, 1 - pf(F, 1, n-2)
8. Or compare the observed F with the critical F, qf(1 - 0.05, 1, n-2).

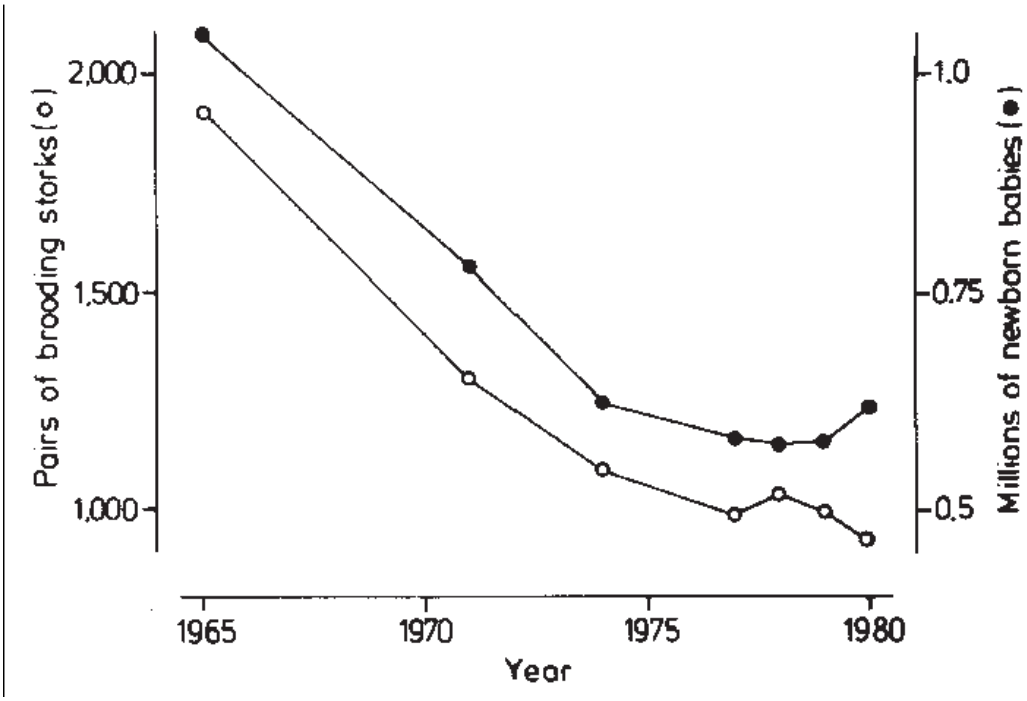
Nonlinear regression

1. Calculate the best-fit curve in R using nls().
2. Plot the best fit curve over the scatter diagram using curve().
3. Estimate the standard error of each parameter in the equation.
4. Form the t value and test for the significance of each parameter in the equation.
5. Or let R do it all with summary(nls()).

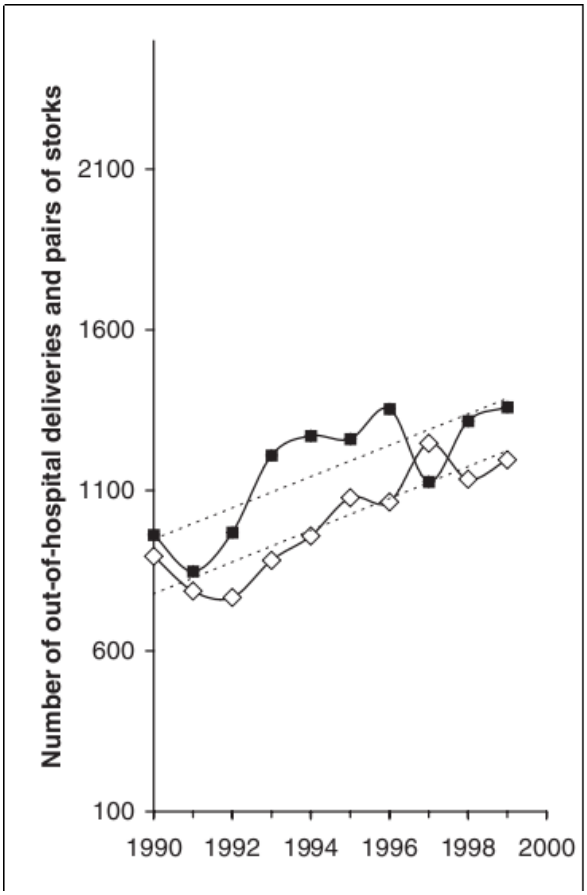
Assumptions of linear regression

1. The x values are measured without any error
2. The relationship between x and y is linear
3. The residuals are normally distributed
4. The variance is constant (regardless of the value of x)

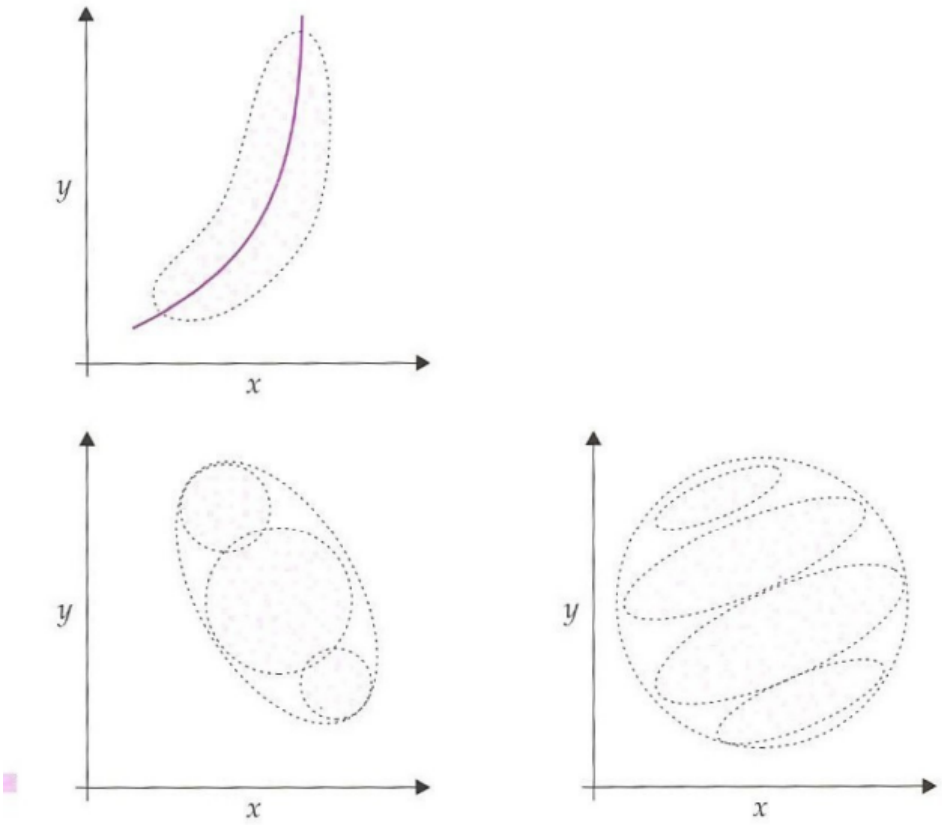
Correlation versus causation



Correlation versus causation



Interpretation of correlation



Lecture 25: Checking assumptions, non-parametric tests

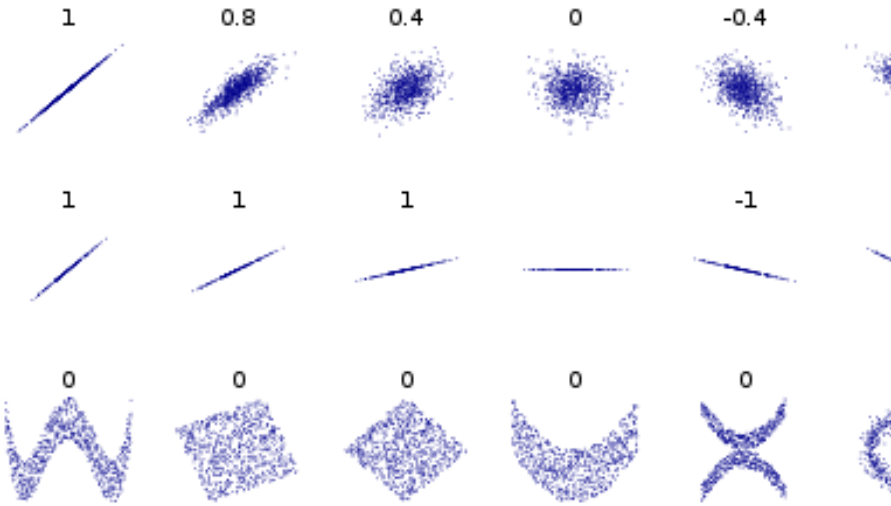
Announcements:

- Reading: Chapter 7 in Vasilj.
- Midterm Exam 3 on Tuesday, January 24, 14:00 to 18:00.
- Today: Checking assumptions of parametric tests in R, non-parametric tests

Correlation versus regression

- Regression: we want to predict y from knowing x
- Correlation: we don't want to predict, just know if there is a relationship
- The probability of the significance test is the same in both
- Regression: calculate the best fit line, ANOVA table
- Correlation: calculate the correlation coefficient

Correlation coefficient



Correlation coefficient $= r = \frac{Cov(x,y)}{s_x s_y}$

Correlation coefficient

korelacijski koeficijent (r)	jačina korelacije
0.00 – 0.10	nema
0.10 – 0.25	vrlo slaba
0.25 – 0.40	slaba
0.40 – 0.50	srednja
0.50 – 0.75	jaka
0.75 – 0.90	vrlo jaka
0.90 – 1.00	potpuna

Checking assumptions

Review: Least Squares Statistics

- Total variation = Explained variation + Unexplained variation
- Sums of squares: $SS_T = SS_{exp} + SS_{res}$
- Degrees of freedom: $df_T = df_{exp} + df_{res}$
- Mean squares: $MS = SS/df$
- F ratio: $F = MS_{exp}/MS_{res}$
- Probability: P = 1 - pf(F, dfexp, dfres)
- F_{crit} = qf(1 - 0.05, dfexp, dfres)

Lecture 25: Checking assumptions, non-parametric tests

- All least squares statistics can be expressed as an ANOVA table.
- In the ANOVA table, variation is partitioned into explained and residual.
- If the ratio of explained to residual is high, then we reject the null hypothesis.
- Null hypothesis is that the predictor variable is not related to the response variable.
- This hypothesis test has several assumptions that we must check.

Assumptions of least squares statistical tests

1. The model is correct.
2. Predictor variables are measured exactly.
3. Residuals are independent of each other.
4. Residuals are independent of predictor variable and response variable.
5. Residuals are normally distributed.
6. Residuals have constant variance.

We can test these assumptions

- Correct model: residuals should be independent of predictor
- Exact measurement of predictor: can't easily test
- Residuals independent: runs test (we won't do this)
- Residuals independent of other variables: runs test, correlation with other variables
- Residuals normally distributed: shapiro.test(residuals), qnorm(residuals), plot(lm(y ~ x))
- Residuals have constant variance: bartlett.test(y x), correlation with other variables, plot(lm(y ~ x))

If the data do not meet the assumptions

Two choices:

1. Transform the response variable (for example, log, square root) and retest the data.
2. Use a non-parametric test instead.
3. Note: the null hypothesis is now different!

Non-parametric tests

1. Non-parametric single factor ANOVA: Kruskal-Wallis test, kruskal.test()
2. Non-parameteric two factor ANOVA: Friedman test, friedman.test()
3. Non-parametric multiple comparisons: Pairwise Wilcoxon test, pairwise.wilcox.test()
4. Or, just rank transform and do a parametric test.
5. Null hypothesis concerns medians rather than means

Assumption of non-parametric tests

1. Residuals are independent of each other
2. Probability distribution is the same for each residual

Non-parametric tests do not assume that residuals are normally distributed