STAT:3510 (22S:101) Biostatistics

Dale Zimmerman

Summer 2016

Definition of biostatistics

- Statistics — the science of collecting, describing, analyzing, and interpreting data, so that inferences (conclusions about a population based on data from merely a sample) can be made with quantifiable certainty.

- Biostatistics — that portion of statistics that is most relevant to the biological sciences.

The definition implies that Statistics requires defining a population of interest, drawing a sample from that population, measuring something on each member of that sample, making a conclusion about the population based on the measured quantities and, finally, using probability to say something about how sure you are of your conclusion.

Example applications of Biostatistics

We use biostatistics to address questions like:

- Does drinking a glass of red wine every day reduce a person's risk of heart disease?
- Does listening to classical music while pregnant increase the musical ability and/or the intelligence of the child in the womb?
- Do larger male dragonflies defend larger territories?
- Does the presence of wind turbines decrease bird populations?

Data collection

Some issues:

- The sample drawn from the population should, if possible, be a *random sample*, i.e. drawn in such a way that every member of the population has an equal chance of being included in the sample.
- Sample size, n (larger n is better, all other things being equal)
- Accuracy of measurement

For the most part, we will not do our own data collection in this class, but will use existing data sets.

Descriptive Statistics

Summaries or reductions of the data into something easier to understand. May be:

- numerical
- tabular
- graphical

Being a summary, a descriptive statistic may reveal some important feature(s) of the data, but not others.

Data Types

The kinds of descriptive statistics that are most appropriate depend on the *type* of data we have collected.

We consider four data types:

- Continuous — numbers that can take on any value in an interval
- Discrete — numbers which are restricted to "isolated" values
- Ordinal (ranked) — ordered labels or categories
- Categorical (nominal) — unordered labels or categories

Data types: Continuous data

Examples:

- Weights of newborn infants
- Elapsed time from swine flu exposure to first symptoms
- Blood lead concentrations in kindergarten children

The data arise by measurement. All arithmetic operations (addition, subtraction, multiplication, division) on the data are meaningful.

Data types: Discrete data

Examples:

- Number of children in family
- Number of gold medals won by U.S. in Winter Olympics
- Number of lung cancer cases for each county in U.S. in 2014

The data arise by counting. All arithmetic operations are meaningful, but proper interpretations require common sense (e.g. an average of 1.73 children per family).

Data types: Ordinal data

Examples:

- Burn severity (1st, 2nd, 3rd degree burns)
- Opinion questionaire items (strongly agree, somewhat agree, neither agree nor disagree, somewhat disagree, strongly disagree)

Can assign numbers to the category labels; ranking these numbers is meaningful, but reporting results of arithmetic operations on them is questionable.

Data types: Categorical data

Examples:

- Gender (male or female)
- Ethnicity (white, black, Asian, . . .)
- Blood type (A+, A-, B+, B-, O+, O-, AB+, AB-)

Can assign numbers to the category labels, but the order and magnitude of the numbers have no meaning. So almost no mathematical operations can be performed on the data (exception: counting the number of individuals that fall into each category).

Data types: Final remarks

- The line between continuous and discrete data may sometimes appear blurry, due to measurement devices which are not "infinitely accurate." Key discriminator: Would the data be discrete if we could measure to an infinite level of accuracy?

- Generally as we move from continuous $\rightarrow$ discrete $\rightarrow$ ordinal $\rightarrow$ categorical, the data become cruder and less informative. Data collection may require trade-offs (larger sample size versus each datum being more informative).

Descriptive statistics: Measures of Center

One important characteristic of a set of data is its "center," although there are different ways to define center.

Three common measures of center:

- Mean
- Median
- Mode

Measures of Center: The Mean

If we represent the numbers in our data set generically as

$$X_1, X_2, \ldots, X_n$$

then their mean, $\overline{X}$ (read "X bar"), is

$$\overline{X} = (X_1 + X_2 + \cdots + X_n)/n.$$

More compactly, using summation notation,

$$\overline{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Measures of Center: The Mean

Clearly, in order to compute the mean and have it be meaningful, the data must be either discrete or continuous.

Toy examples:

- Data 1,2,3,4,5: $\overline{X} = (1+2+3+4+5)/5 = 3$
- Data 6,7,8,9,10: $\overline{X} = (6+7+8+9+10)/5 = 8$
- Data 4,6,8,10,12: $\overline{X} = (4+6+8+10+12)/5 = 8$
- Data 1,1,1,1,36: $\overline{X} = (1+1+1+1+36)/5 = 8$
- Data 2,3,6,8,11,18: $\overline{X} = (2+3+6+8+11+18)/6 = 8$

Measures of Center: The Mean

The mean is the "balance point" for the data, i.e. it is where a fulcrum would need to be put to balance equally weighted objects placed on the number line at $X_1, X_2, \ldots, X_n$.

Equivalently, $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$.

Measures of Center: The Median

The median is the middle value in the ordering of all data values from smallest to largest. Clearly this requires the data to be ordinal, discrete, or continuous.

If we represent the ordered values in our data set generically as

$$X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$$

then their median, $\tilde{X}$ (read "X tilde"), is

$$\tilde{X} = \begin{cases} X_{(\frac{n+1}{2})} & \text{if } n \text{ is odd} \\ (X_{(\frac{n}{2})} + X_{(\frac{n}{2}+1)})/2 & \text{if } n \text{ is even} \end{cases}$$

Measures of Center: The Median

Toy examples:

- Data 1,2,3,4,5: $\tilde{X} = 3$
- Data 6,7,8,9,10: $\tilde{X} = 8$
- Data 4,6,8,10,12: $\tilde{X} = 8$
- Data 1,1,1,1,36: $\tilde{X} = 1$
- Data 2,3,6,8,11,18: $\tilde{X} = (6+8)/2 = 7$

The median is also known as the 50th percentile.

Measures of Center: The Mode

The mode is the datum that occurs most frequently in the sample.

Toy examples:

- Data 1,1,1,1,36: Mode= 1
- Data 2,3,6,8,11,18: No mode (or, alternatively, every datum is a mode)
- Data 2,3,3,8,11,18,18: Two modes (bimodal), 3 and 18

The mode is the most common value, but it may or may not be representative of the dataset's center.

Mean vs. Median vs. Mode

- Mode is well-defined for all data types, median requires at least ordinal data, mean requires at least discrete data.
- Mode often useless for continuous data.
- Mode is a datum; median may or may not be a datum (depending on whether n is odd or even); mean often is not a datum.
- Mean has superior statistical properties (to be seen later).
- Mean is distorted more than the others if the data are skewed (definition to come later) or contain outliers.
- Units for all three are the same as the units of the data.

Descriptive Statistics: Measures of Dispersion (Spread)

Another important attribute of a dataset is how spread out it is from its center.

Example: Dataset #1 (6,7,8,9,10) versus Dataset #2 (0,4,8,12,16).

Practical applications: pill dosage, nuts and bolts

Measures of Dispersion: The Range

$$\text{Range} = X_{(n)} - X_{(1)}$$

Features:

- Very easy to compute
- Because it's based on only 2 values, it is very sensitive to outliers
- Because it's based on only 2 values, it does not reflect the variability in the data that lie between the two extremes

Measures of Dispersion: The Interquartile Range (IQR)

$$\text{IQR} = Q_3 - Q_1$$

where Q_1 and Q_3 are the first and third quartiles (Q_2, the second quartile, coincides with the median).

How are the first and third quartiles defined?

- Q_1 is the median of the observations less than Q_2 (i.e., Q_1 is the median of the lower half of the ordered sample)
- Q_3 is the median of the observations greater than Q_2 (i.e., Q_3 is the median of the upper half of the ordered sample)

Measures of Dispersion: The Interquartile Range (IQR)

Features of the IQR:

- Not as easy to compute as the range
- Much less sensitive to outliers than the range
- Still, it's not as informative as a measure that would utilize all the data

Measures of Dispersion: The Variance

We'd like a measure of spread that utilizes information from all the observations. How about the mean deviation, $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})$?

We can avoid the problem of negative deviations canceling out positive deviations by squaring each deviation, i.e. the mean squared deviation

$$\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

The variance, s^2, is very similar to this:

$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

(Using a divisor of $n-1$ rather than n improves some statistical properties.)

Measures of Dispersion: The Variance

There is an alternate formula for s^2, called the "computational formula," which is usually easier to calculate (and less susceptible to careless errors) than the one on the previous page:

$$s^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}\right)$$

Note that this involves summing the squares of the data (i.e. to compute $\sum_{i=1}^{n} X_i^2$ we square first and then sum), as well as squaring the sum of the data.

Measures of Dispersion: The Variance

Features of the variance:

- Units of s^2 are the squares of the units of observation. To convert back to the units of observation, we take the (positive) square root, yielding the **standard deviation**, s:

$$s = \sqrt{s^2}$$

- s^2 (and s) is affected by outliers less than the range, but more so than IQR.

Toy examples

- Data 6,7,8,9,10: Range $= 10-6 = 4$, IQR $= 9-7 = 2$,

  $$s^2 = \frac{1}{4}[(6-8)^2+(7-8)^2+(8-8)^2+(9-8)^2+(10-8)^2] = 2.5,$$

  $$s^2 = \frac{1}{4}[(6^2+7^2+8^2+9^2+10^2)-\frac{40^2}{5}] = \frac{1}{4}(330-320) = 2.5,$$

  $s = \sqrt{2.5} = 1.58$

- Data 4,6,8,10,12: Range $= 12-4 = 8$, IQR $= 10-6 = 4$, $s^2 = 10$, $s = 3.16$

- Data 1,1,1,1,36: Range $= 36-1 = 35$, IQR $= 1-1 = 0$, $s^2 = 245$, $s = 15.65$

Sample statistics vs. population parameters

Consider taking a sample (hopefully random) from a population and recording some variable for each member of the sample. When computed from this sample, the measures of center and spread that we've discussed are called *statistics* and are often called the sample mean, sample median, . . . , sample standard deviation.

Conceptually, we can imagine computing the same measures for the entire population; yielding the population mean, population median, etc. These measures are called *parameters*.

Each statistic can be thought of as an estimate of the corresponding parameter. E.g., the sample variance is an estimate of the population variance. As n increases, the estimate tends to get closer to the corresponding parameter.

Sample statistics vs. population parameters

In some rare instances the population is sufficiently small and accessible that we can record the variable for every member of the population. In such cases we can calculate the mean and variance of the population — we don't need to estimate it. These calculations for the population are performed just as for the sample, except that a divisor of N (the population size) is used in place of $n - 1$ in the calculation of the population variance.

Computing $\overline{X}$ and s^2 from grouped data

Some discrete, ordinal, or categorical datasets are large, but if the number of distinct values in these datasets is small they can be represented compactly by a frequency table.

Example (from Problem 5 in Chapter 1): 100 sampling quadrats were surveyed in NY, $X_i =$ # of *Cepaea nemoralis* snails per quadrat.

# of snails, $X_{[i]}$	Frequency, f_i
0	69
1	18
2	7
3	2
4	1
5	1
8	1
15	1
	100

Computing $\overline{X}$ and s^2 from grouped data

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i = \frac{1}{n}\sum_i f_i X_{[i]},$$
$$s^2 = \frac{1}{n-1}\left(\sum_i f_i X_{[i]}^2 - \frac{(\sum_i f_i X_{[i]})^2}{n}\right)$$

# of snails, $X_{[i]}$	f_i	$f_i X_{[i]}$	$f_i X_{[i]}^2$
0	69	0	0
1	18	18	18
2	7	14	28
3	2	6	18
4	1	4	16
5	1	5	25
8	1	8	64
15	1	15	225
	100	70	394

$$\overline{X} = 70/100 = 0.7, \quad s^2 = \frac{394 - \frac{70^2}{100}}{99} = 3.48, \quad s = 1.87$$

Linear transformations of data

Sometimes we may have computed a measure of center or spread for data recorded in one type of units (e.g. minutes, pounds), but we want to have instead the corresponding measure of center or spread for the data recorded in different units (seconds, kilograms).

One option is to transform every datum to the new units, and recompute the measure of center or spread.

But this is unnecessarily tedious if the new units are a *linear transformation* of the original units. A linear transformation of data $X_1, X_2, \ldots$ yields transformed data $Y_1, Y_2, \ldots, Y_n$ via the equation

$$Y_i = aX_i + b, \qquad i = 1, 2, \ldots, n.$$

Linear transformations of data

Examples:

- Minutes $\rightarrow$ seconds: $Y_i = 60X_i + 0$
- Pounds $\rightarrow$ kilograms: $Y_i \doteq \frac{1}{2.20462262} X_i + 0$
- °F $\rightarrow$ °C: $Y_i = \frac{5}{9}(X_i - 32) = \frac{5}{9}X_i - \frac{160}{9}$

Note: not all transformations are linear. One example of a nonlinear transformation is the logarithmic ("log") transformation, $Y_i = \log X_i$.

Linear transformations of data

What happens to measures of center when the data are linearly transformed? Consider the following examples:

- Original data (1,2,3,4,4,4): $\overline{X} = 3$, $\tilde{X} = 3.5$, mode $= 4$
- Add 100 to each datum ($Y_i = X_i + 100$), yielding transformed data (101,102,103,104,104,104): $\overline{Y} = 103$, $\tilde{Y} = 103.5$, mode $=$ 104
- Multiply each original datum by 5 ($Y_i = 5X_i$), yielding transformed data (5,10,15,20,20,20): $\overline{Y} = 15$, $\tilde{Y} = 17.5$, mode $= 20$

In general, $\overline{Y} = a\overline{X} + b$ (with a similar result for median and mode).

Linear transformations of data

What happens to measures of spread when the data are linearly transformed? Consider the same example:

- Original data (1,2,3,4,4,4): Range$=3$, IQR$=2$, $s_X^2 = 1.6$, $s_X = 1.26$
- Add 100 to each datum ($Y_i = X_i + 100$), yielding transformed data (101,102,103,104,104,104): Range$=3$, IQR$=2$, $s_Y^2 = 1.6$, $s_Y = 1.26$
- Multiply each original datum by 5 ($Y_i = 5X_i$), yielding transformed data (5,10,15,20,20,20): Range$=15$, IQR$=10$, $s_Y^2 = 40$, $s_Y = 6.32$

In general, $s_Y = |a| s_X$ (with a similar result for range and IQR), while $s_Y^2 = a^2 s_X^2$.

Accuracy and Precision

Accuracy is the closeness of a measured or computed value to its true value.

Precision is the closeness of repeated measurements of the same quantity to each other (regardless of whether they are close to the true value).

The # of digits used for recording continuous data implies a certain level of precision. The "30–300 rule" (see next slide) should be used to determine this level. Whatever the level of precision of the data, measures of center and all measures of spread except the variance should be calculated to one additional digit. The variance should be calculated to two additional digits. Example:

Data $1.7, 1.2, 2.9, 2.1$: $\quad \overline{X} = 1.975 \to 1.98,\ s^2 = 0.5158333 \to 0.516.$

The 30–300 Rule

As noted previously, the greater the level of precision in the measured data, generally the more costly (in time and effort) it is to collect that data, as well as to compute numerical measures such as the mean and variance.

But measuring too crudely (e.g. measuring height of people to the nearest meter) may render the entire inferential enterprise worthless.

The 30–300 rule says to measure data to a level of precision for which there are at least 30, but not more than 300, unit steps between the smallest and largest measurements.

Example: If the smallest and largest heights of students in this class are 4’8” and 6’4”, then we should record height to the nearest 0.5”.

Frequency distributions (in tabular form)

Recall the frequency table used to summarize data from 100 sampling quadrats surveyed in NY, X_i =# of *Cepaea nemoralis* snails per quadrat:

# of snails, $X_{[i]}$	Frequency, f_i
0	69
1	18
2	7
3	2
4	1
5	1
8	1
15	1
	100

We can add columns to this table that give *relative frequencies* (proportions of the total # of observations that fall in each category) and *cumulative relative frequencies* (proportions of the total # of obser-

vations that fall in each category or previous categories).

# of snails, $X_{[i]}$	Frequency, f_i	Relative freq	Cum rel freq
0	69	0.69	0.69
1	18	0.18	0.87
2	7	0.07	0.94
3	2	0.02	0.96
4	1	0.01	0.97
5	1	0.01	0.98
8	1	0.01	0.99
15	1	0.01	1.00
	100	1.00	

The example above involves discrete data; we can do a similar thing for continuous data but we have to partition the interval containing the data into several subintervals of equal size and do the counting within those subintervals.

Frequency table for first age guess data

Age guess	Frequency	Relative freq	Cum rel freq
$[30-35)$	1	0.01	0.01
$[35-40)$	2	0.03	0.04
$[40-45)$	4	0.06	0.10
$[45-50)$	26	0.39	0.49
$[50-55)$	28	0.42	0.91
$[55-60)$	6	0.09	1.00
	67	1.00	

Bar graphs

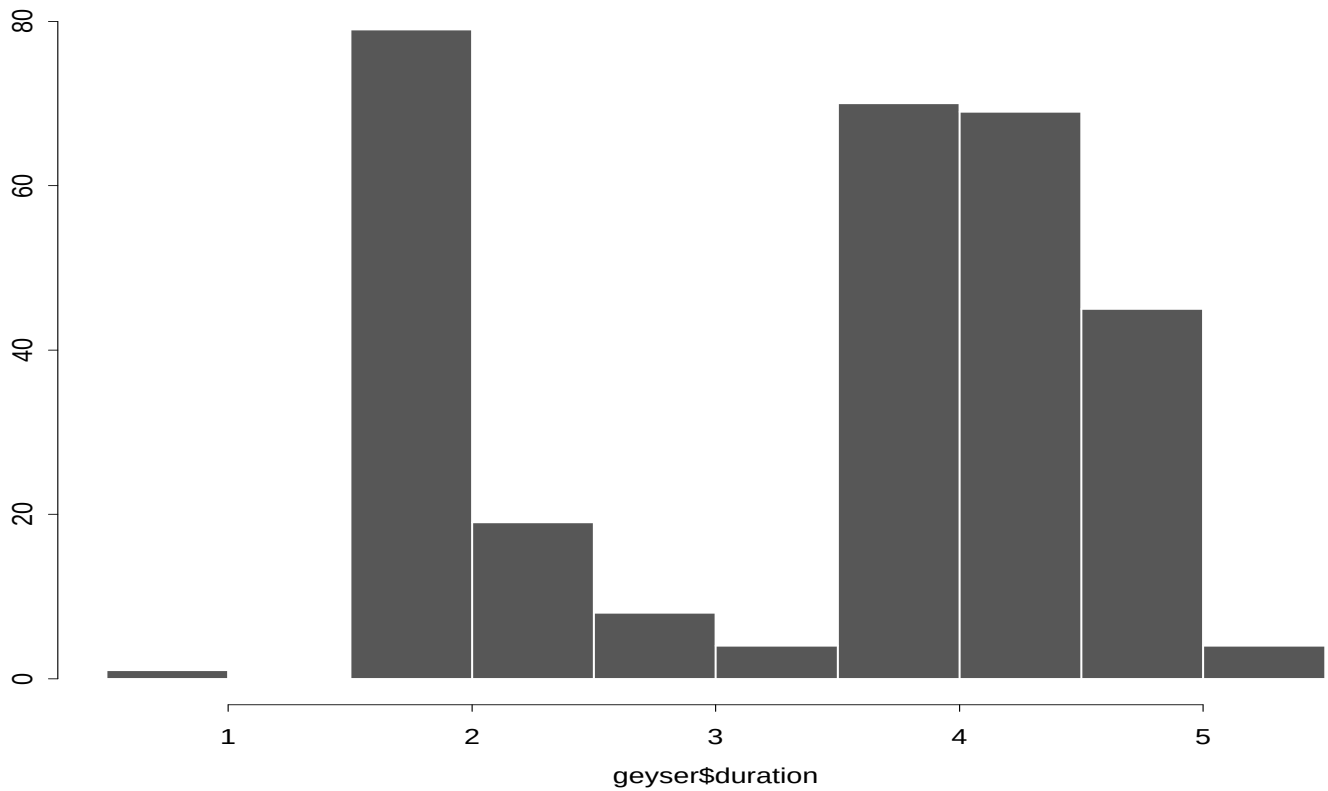
- A graphical display of the distribution of frequencies (or relative frequencies)
- Used for discrete, ordinal, or categorical data
- Simply plot a bar, centered at each data value, whose height is equal to the corresponding (relative) frequency

Bar graph for snail data

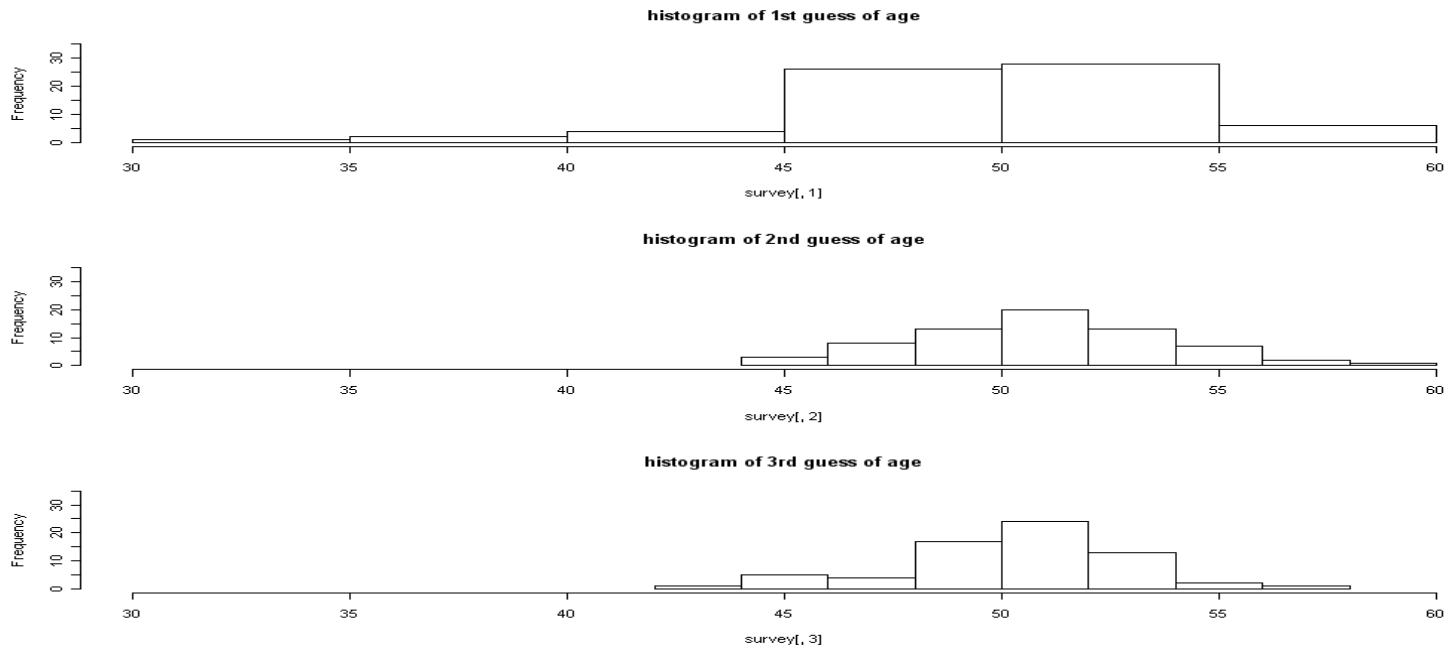
Histograms

- Similar to a bar graph
- Used for continuous data
- Have to choose how many subintervals to use, and where to start the first and end the last
- Generally 5 to 15 subintervals work best: $< 5 \rightarrow$ oversummarization, and $> 15 \rightarrow$ undersummarization.

Histogram of Old Faithful geyser eruption durations



Histogram of guesses of Dr. Z's age



Data shape

In addition to center, spread, and outliers, a bar graph/histogram displays the *shape* of the data.

The data's relative frequency distribution is said to be *symmetric* (approximately) if the bar graph/histogram is symmetric (approximately) around its center.

Examples of symmetric distributions:

Data shape: Skewness

If the data are not symmetric, they may be *skewed*.

- Right skewed — a longer right tail
- Left skewed — a longer left tail

Examples of skewed distributions:

Effects of data shape on measures of center

- For a symmetric distribution, mean = median
- For a symmetric unimodal distribution, mean = median = mode
- For a right skewed distribution, mode < median < mean
- For a left skewed distribution, mean < median < mode

Five-number summaries and box plots

The *five-number summary* of a dataset consists of the numbers in the following list (and in this order):

$$\text{Minimum, } Q_1\text{, Median, } Q_3\text{, Maximum}$$

A *box plot* of a dataset is a graphical version of the five-number summary, with a few extras.

Generic box plot:

Step-by-step procedure for constructing a box plot

1. Draw a horizontal (or vertical) reference scale based on the extent of the data.

2. Draw a box whose sides (or top and bottom) are located at Q_1 and Q_3.

3. Draw a vertical (horizontal) line segment at the median.

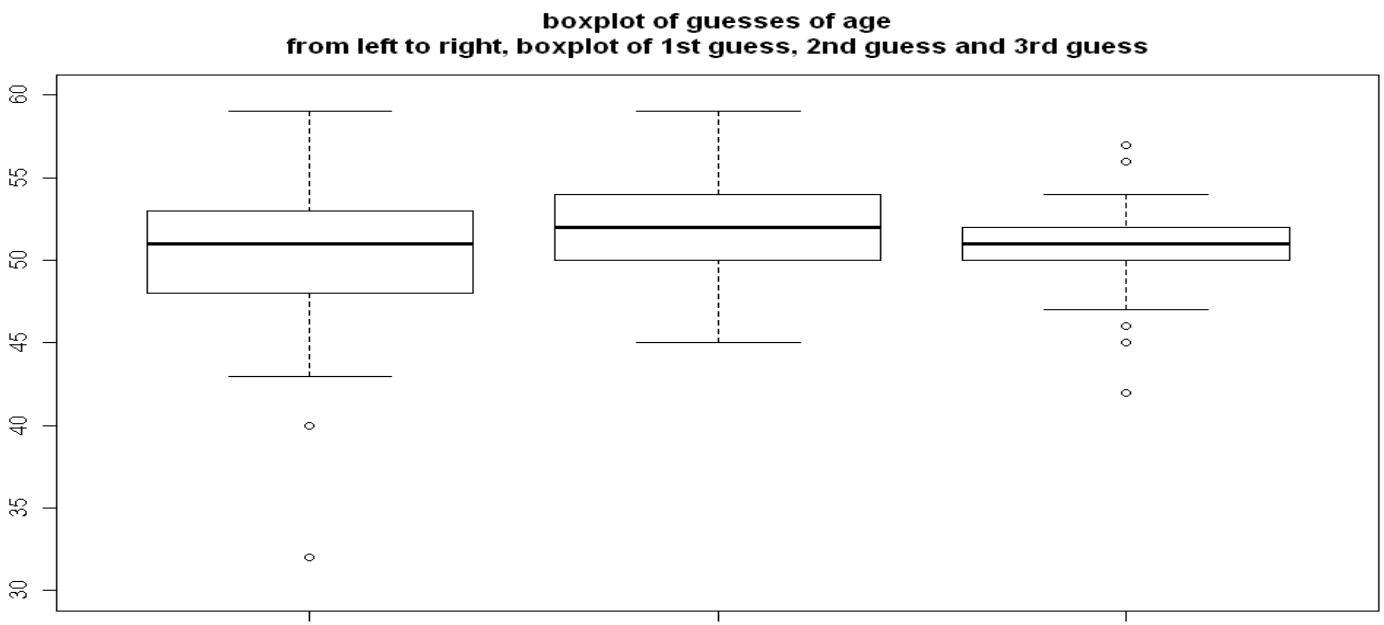
4. Compute the *fences*, $f_1 = Q_1 - 1.5 * IQR$ and $f_3 = Q_3 + 1.5 * IQR$.

5. Extend a line segment (so-called *whisker*) from Q_1 out to the most extreme observation that is at or inside the fence, i.e., $\geq f_1$). Repeat on the other side, i.e., from Q_3 to the most extreme observation that is $\leq f_3$. Mark the end of these line segments with a $\times$.

6. Mark any observations beyond the fences with an open circle, ○; these are regarded as outliers.

7. If you are constructing more than one box plot for comparison purposes, use the same scale for all of them and put them side-by-side (or one on top of another)

Box plots for guesses of Dr. Z's age



Describing shape from a box plot

- For an (approximately) symmetric distribution, Q_2 will be near the middle of the box, and the two whiskers will be nearly the same length.

- Right (left) skewness lengthens the box and the whisker to the right (left) of the median, relative to the lengths of the box and whisker on the other side of the median.

Probability: Basic concepts and terminology

Probability is a number between 0 and 1 (inclusive) that measures how likely something is to occur. A more formal definition will be given subsequently.

An *experiment* is an activity with an outcome that is observable but not predictable with certainty (so it's sometimes called a *random experiment*). Examples:

- Rolling a six-sided die
- Taking an aspirin when you have a headache and determining whether it relieved the pain
- Randomly sampling an incoming UI freshman and measuring their HS GPA

Probability: Basic concepts and terminology

The *sample space* is the collection of all possible outcomes of the experiment. Examples from previous 3 experiments:

- $S = \{1, 2, 3, 4, 5, 6\}$
- $S = \{\text{pain relieved, pain not relieved}\}$
- $S = [0.00, 4.00]$

Probability: Basic concepts and terminology

An *event* is a subset of the sample space, i.e. an outcome or a subset of outcomes of the experiment. Usually represent events using symbols $A, B, C, \ldots$. Examples from previous 3 sample spaces:

- $A = \{1, 2\}$, $B = \{2, 4, 6\}$
- $A = \{\text{pain not relieved}\}$
- $A = 2.93$, $B = [2.5, 3.0)$

Probability, P, is a mathematical function that assigns a unique number between 0 and 1 (inclusive) to every possible event of an experiment. It must obey certain axioms (see p. 35 of text).

We write the probability that event A occurs as $P(A)$.

Probability: Basic concepts and terminology

If a sample space consists of a finite #, say n, of outcomes, and the outcomes are equally likely, then the axioms of probability can be used to show that each outcome has probability $1/n$ and any event with k outcomes has probability k/n.

Example: Consider rolling a fair six-sided die. The sample space is $S = \{1,2,3,4,5,6\}$ and each of the six outcomes is equally likely. So, for instance, if we define $A = \{2,4,5\}$, then $P(A) = \frac{3}{6} = \frac{1}{2}$.

The definition of probability based on the aforementioned axioms is known as *classical probability*.

Probability: Basic concepts and terminology

Another type, or interpretation, of probability is the *relative frequency* definition. Under this definition, we consider repeating the experiment a large number, N of times. The relative frequency probability of an event A is given by

$$\frac{\#\text{ times } A \text{ occurs}}{N}.$$

When the classical probability of A is known, it turns out that it coincides with

$$\lim_{N\to\infty} \frac{\#\text{ times } A \text{ occurs}}{N}.$$

Counting outcomes

Sometimes the sample space of an experiment consists of n equally likely outcomes but n is not so easy to determine. This is particularly so when the experiment is a *multi-stage* or *composite* experiment, i.e. an experiment built up from smaller experiments. For those cases we need to learn some rules about counting outcomes.

Example 1: Toss a fair coin 6 times. This experiment is made up of 6 smaller experiments, each of which is to toss a fair coin once. Each smaller experiment has 2 possible (and equally likely) outcomes: H for "heads" and T for "tails". How many outcomes does the composite experiment have?

Counting outcomes

The **Multiplication Rule** says that the number of outcomes for a composite experiment is obtained by multiplying the number of outcomes for the individual experiments together, subject to any imposed restrictions on repetition (or any other restrictions).

So for Example 1, the number of outcomes for the composite experiment is

$$2 \times 2 \times 2 \times 2 \times 2 \times 2 = 2^6 = 64.$$

Thus, $P(H,H,T,T,H,T) = \frac{1}{64}$,
and $P(\text{first and last tosses are heads}) = \frac{2^4}{64} = \frac{1}{4}$.

In general, if a composite experiment consists of k smaller experiments, each with n possible outcomes, and repetition of outcomes is allowed, then the composite experiment has n^k possible outcomes.

Counting outcomes

Example 2: There are four different stimuli to be applied to crayfish. Four crayfish are available. Each stimulus is to be applied to one and only one crayfish. How many different assignments of stimulus to the crayfish are possible?

Answer: There are 4 choices of stimulus for the first crayfish, 3 choices for the second, 2 for the third, and 1 for the last crayfish. Thus the number of different assignments of stimulus to the 4 crayfish is

$$4 \times 3 \times 2 \times 1 = 4! = 24.$$

In general, if a composite experiment consists of k smaller experiments, each with n possible outcomes, and repetition of outcomes is not allowed, then the composite experiment has $n \times (n-1) \times \cdots \times (n-k+1) = \frac{n!}{(n-k)!}$ possible outcomes.

Counting outcomes

Example 3: There are four different stimuli that can be applied to crayfish. Two of the four stimuli are to be randomly selected and applied to one crayfish (i.e. the crayfish receives both stimuli), and the order in which it receives the two stimuli is regarded as irrelevant. How many different stimulus pairs are there?

Answer: 6; specifically they are $\{(s_1, s_2), (s_1, s_3), (s_1, s_4), (s_2, s_3), (s_2, s_4), (s_3, s_4)\}$. This is the # of ways of choosing 2 stimuli from 4 stimuli, without regard to order.

In general, the # of ways of choosing k items from n items, without regard to order, is

$$\frac{n!}{k!(n-k)!} \equiv \left(\begin{array}{c} n \\ k \end{array} \right).$$

Counting outcomes

Example 3, continued: Thus, the probability that s_3 is among the two stimuli selected is calculated as

$$\frac{\#\text{ of outcomes in } \{(s_1, s_3), (s_2, s_3), (s_3, s_4)\}}{\binom{4}{2}} = \frac{3}{6} = \frac{1}{2}.$$

Example 4: Problem 2.3(a).

Complement of an event

Let A be an event. The *complement* of A, denoted A', consists of all outcomes of the experiment that are not in A.

Complement in a Venn diagram:

Example 1: Roll a fair die once. If $A = \{3,6\}$, then $A' = \{1,2,4,5\}$.

Example 2: Toss a fair coin 6 times. If A=$\{$first and last tosses are heads$\}$, then A'=$\{$either the first toss or the last toss (or both) are tails$\}$.

Probability of complement:

$$P(A') = 1 - P(A).$$

Intersection of two events

Let A and B be two events. The *intersection* of A and B, denoted by $A \cap B$, consists of all outcomes that are in both A and B.

Intersection in a Venn diagram:

Example: Roll a fair die once. If $A = \{3,6\}$, $B = \{2,3,4\}$, and $C = \{1,2,4\}$, then:

$$\begin{aligned} A \cap B &= \{3\}, \\ A \cap C &= \emptyset, \\ B \cap C &= \{2,4\}. \end{aligned}$$

Conditional probability

Let A and B be two events. The *conditional probability*, $P(A|B)$, is the probability that A occurs given that B has occurred.

The probability of the intersection of two events is related to the conditional probability of one event given the other through the *multiplicative rule of probability*:

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

Provided that $P(A) \neq 0$, we may re-express the first of these equalities as follows (with a similar result for the second equality):

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

Conditional probability examples

In the die roll experiment, if $A = \{1,2\}$, $B = \{1,3,5\}$, and $C = \{1,2,3\}$, then:

Independent events

Events A and B are said to be *independent* if the occurrence of one of them has no effect on the probability of occurrence of the other, i.e. if

$$P(A|B) = P(A) \quad \text{or equivalently} \quad P(B|A) = P(B).$$

As a consequence of the multiplicative rule of probability, if A and B are independent then the probability of their intersection is

$$P(A \cap B) = P(A) \cdot P(B).$$

Example: Roll a fair die once. Let $A = \{1,2\}$, $B = \{1,3,5\}$, $C = \{1,2,3\}$. Then A and B are independent, but A and C are not independent.

Mutually exclusive events

Events A and B are said to be *mutually exclusive* (or *disjoint*) if they have no outcomes in common, i.e. if their intersection is empty.

Mutually exclusive events in a Venn diagram:

Example: Roll a fair die once. Then $A = \{3,6\}$ and $B = \{1,4,5\}$ are mutually exclusive.

Note: If A and B are mutually exclusive, then $P(A \cap B) = 0$ (and vice versa).

Union of two events

Let A and B be two events. The *union* of A and B, denoted by $A \cup B$, consists of all outcomes that are in either A or B (or both).

Union in a Venn diagram:

Example: Roll a fair die once. If $A = \{3,6\}$, $B = \{2,3,4\}$, and $C = \{1,2,4\}$, then:

$$\begin{aligned} A \cup B &= \{2,3,4,6\}, \\ A \cup C &= \{1,2,3,4,6\}, \\ B \cup C &= \{1,2,3,4\}. \end{aligned}$$

Probability of the union of two events

In general,

$$P(A \cup B) = P(A) + P(B) - P(A \cap B).$$

However, if A and B are mutually exclusive, this equality simplifies to

$$P(A \cup B) = P(A) + P(B).$$

Applying these to the example on the previous slide, we obtain

$$\begin{aligned} P(A \cup B) &= \frac{1}{3} + \frac{1}{2} - \frac{1}{6} = \frac{2}{3} \\ P(A \cup C) &= \frac{1}{3} + \frac{1}{2} - 0 = \frac{5}{6}. \end{aligned}$$

Probability practice with a biological problem

A study in the pinyon pine woodlands of northern Arizona examined the susceptibility of pinyon pine trees to drought-induced mortality. The study area is often exposed to severe drought. Some pinyon pine trees contain a protein called glycerate dehydrogenase (GLY), which is suspected to increase a tree's drought resistance. Suppose that a pinyon pine tree is selected at random from the study area, and that:

$P(D) = 0.20$ where $D = \{\text{tree is dead}\}$,
$P(G) = 0.70$ where $G = \{\text{GLY is present in tree}\}$, and
$P(D \cap G) = 0.07$.

Probability practice with a biological problem (cont'd)

1. Calculate the probability that the selected tree is dead or contains GLY (or both).

Probability practice with a biological problem (cont'd)

2. Calculate the probability that the selected tree is dead, given that it contains GLY.

Probability practice with a biological problem (cont'd)

3. Calculate the probability that the selected tree is dead, given that it does not contain GLY.

Probability practice with a biological problem (cont'd)

4. Do the results of Problems 2 and 3 provide evidence to support the suspicion that GLY provides some resistance to drought? Explain briefly.

Probability practice with a biological problem (cont'd)

5. Are events D and G independent, mutually exclusive, both, or neither? Justify your answer.

Bayes Rule

In some biological settings we know $P(A)$, $P(B)$, and $P(A|B)$, and we want to find $P(B|A)$. We can obtain it using Bayes Rule:

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

If we do not know $P(A)$ but we know $P(A|B')$, we can substitute for $P(A)$ in the above equation using the result

$$\begin{aligned} P(A) &= P(A \cap B) + P(A \cap B') \\ &= P(B)P(A|B) + P(B')P(A|B'). \end{aligned}$$

Application of Bayes rule: Diagnostic screening

Suppose that a nurse tests a person for TB using the "skin test." Define

$$A = \{\text{skin test positive}\}, \quad B = \{\text{person has TB}\}.$$

Suppose further that for the population of interest, $P(B) = .03$, $P(A|B) = .90$, and $P(A|B') = .05$. Then

$P(B|A) =$

Random variables

A *random variable* is a variable whose outcome is determined at least partly by chance. It results from carrying out a (random) experiment.

We focus (for now) on two types of random variables:

- Discrete random variable — the only values the variable can have are discrete. Examples: # heads in 3 tosses of a fair coin, # snails in a randomly chosen quadrat.

- Continuous — the values the variable can have are continuous. Examples: wt of 1st infant born in UIHC in 2015, survival time of a randomly selected stroke patient.

Random variables

We usually represent random variables by capital letters near the end of the alphabet, especially X, Y, and Z; and we represent a generic value of the variable by the corresponding lower-case letter.

Discrete random variables: the probability density function

For every discrete random variable, the numerical values it can take on can be listed (prior to conducting the random experiment) and a probability can be associated with each one. The function, $f(x) = P(X = x)$, that assigns these probabilities is called the *probability density function (pdf)*.

Example: Toss a fair coin three times, let $X =$# heads. The pdf is:

x	$f(x)$
0	.125
1	.375
2	.375
3	.125

Discrete random variables: the probability density function

More generally, the pdf, f, of a discrete random variable satisfies:

- f is defined for all real numbers;
- $f(x) \geq 0$ (since it's a probability);
- $f(x) = 0$ for most real numbers since X is discrete;
- $\sum_{\text{all } x} f(x) = 1$.

Furthermore, if A is an event, we have

$$P(A) = P(X \in A) = \sum_{\text{all } x \in A} f(x).$$

Discrete random variables: the probability density function

Example: If A in the previous example is $\{$at least one head$\}$, we find that

$$P(A) = P(X=1) + P(X=2) + P(X=3) = .375 + .375 + .125 = .875.$$

If $B = \{$an odd number of heads$\}$, then

$$P(B) = P(X=1) + P(X=3) = .375 + .125 = .500.$$

Discrete random variables: the probability density function

Note: We may display the pdf of a discrete random variable graphically, by something similar to a bar graph for discrete data. For the previous example, we have:

Discrete random variables: the cumulative distribution function

If we apply the same accumulating procedure to the pdf of a discrete variable as we applied to the relative frequencies of a discrete frequency distribution to get the cumulative relative frequencies, we obtain the cumulative distribution function (CDF), $F(x) = P(X \leq x)$.

Example: Toss a fair coin three times, let $X =$ # heads. The CDF is:

x	$f(x)$	$F(x)$
0	.125	.125
1	.375	.500
2	.375	.875
3	.125	1.000

Discrete random variables: the expected value

Suppose we repeated the experiment of 3 tosses of a fair coin a very large number of times. What would be the average value of X over all repetitions of the experiment? Using the pdf, we can determine this analytically rather than empirically.

The *expected value* (long-run average value) of a discrete random variable X is given by

$$\mu = E(X) = \sum_{\text{all } x} xf(x).$$

Example 1: For the experiment of 3 tosses of a fair coin, with $X =$ # of heads, we have

$$E(X) = (0)(.125) + (1)(.375) + (2)(.375) + (3)(.125) = 1.5.$$

Discrete random variables: the expected value

Example 2: The proportions of the 114.4 million U.S. households of various sizes in 2006 were as follows (Source: U.S. Census Bureau):

Household size (x)	Proportion ($f(x)$)
1	.270
2	.330
3	.170
4	.140
5	.060
6	.020
7+	.010

If we randomly sample one household from this population, what is its expected size? Answer: $E(X) = (1)(.270) + (2)(.330) + \cdots + (7)(.010) = 2.49$ (actually 2.49+).

Discrete random variables: the variance

Again, suppose we repeat a random experiment a very large number of times, and observe the outcome of the random variable X for each such experiment. How spread out would we expect these outcomes to be? The answer is provided by the *variance* of X, computed from the pdf as follows:

$$\sigma^2 = E[(X-\mu)^2] = \sum_{\text{all } x} (x-\mu)^2 f(x) = \sum_{\text{all } x} (x^2 f(x)) - \mu^2.$$

Example: For the experiment of 3 tosses of a fair coin, with $X = \#$ of heads, we have

$$\sigma^2 = (0)^2(.125)+(1)^2(.375)+(2)^2(.375)+(3)^2(.125)-1.5^2 = 0.75.$$

The binomial distribution: basic framework

In many biological applications, there is some basic trial for which there are only two possible outcomes, labelled "success" and "failure", and we are interested in the # of successes that occur when we repeat the basic trial n times.

Examples:

- # of heads in 3 tosses of a fair coin
- # of left-handed people in a class of 25 kindergarteners
- # of cancer patients in a clinical trial whose cancer goes into remission

All these are examples of *binomial* random variables.

The binomial distribution: basic framework

More precisely, if:

- A fixed number n of trials are carried out;
- The outcome of each trial is either a "success" or a "failure";
- The probability of success, denoted by p, is constant from trial to trial;
- The trials are independent (the outcome on any particular trial does not affect the outcome of any other trial);

then $X = \#$ successes in the n trials
is a binomial random variable with parameters n and p.

The binomial distribution: the pdf

The pdf of a binomial random variable (with parameters n and p) is called the binomial pdf (with parameters n and p). It turns out that this pdf is given by

$$f(x) = \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x} \quad \text{for } x = 0, 1, 2, \ldots, n.$$

(See text, p. 72 for rationale.)

Example 1: $X =$ # of heads in 3 tosses of a fair coin:

$$f(x) = \frac{3!}{x!(3-x)!}(0.5)^x(1-0.5)^{3-x} = \frac{6}{x!(3-x)!}(0.5)^3 = \frac{3/4}{x!(3-x)!}$$

for $x = 0, 1, 2, 3$. (Check to make sure this matches what we had previously on p. 81 of these notes.)

The binomial distribution: biological examples

Example 2: The proportion of U.S. residents that are lactose intolerant is believed to be about 0.10. If this is correct and we randomly select 7 people from the U.S. population, what is the probability that there are no lactose intolerant people in the sample?

Answer:

The binomial distribution: biological examples

Example 3: The proportion of U.S. residents that have Type A blood is about 40%. If we randomly select 20 people from the U.S. population, what is the probability that at least 40% of the people in the sample have Type A blood?

Answer:

The binomial distribution: the CDF and Table C.1

If X has a binomial distribution with parameters n and p, and if we let d represent any integer between 0 and n, then the CDF of X is given by

$$F(d) = P(X \leq d) = \sum_{x=0}^{d} \frac{n!}{x!(n-x)!} p^x (1-p)^{n-x}.$$

For selected values of n and p, this CDF is given in Table C.1 of our text, and we can use it to avoid having to compute the pdf.

Re-do of Example 1:

The binomial distribution: the CDF and Table C.1

Re-do of Example 2:

Re-do of Example 3:

The binomial distribution: Mean, variance, and shape

If X is a binomial random variable with parameters n and p, then:

- $\mu = E(X) = np$
- $\sigma^2 = np(1-p)$
- the pdf is symmetric if $p = 0.5$; right skewed if $p < 0.5$; and left skewed if $p > 0.5$

The Poisson distribution: basic framework

Another important experimental framework in biology is one in which counts are made of the # of occurrences of some phenomenon (a basic event) in a fixed interval of time or a fixed unit of length, area, or volume.

Examples:

- # of hurricanes in a year
- # of snails in a 1 m $\times$ 1 m quadrat
- # of mutations in a segment of DNA of fixed length

All these are examples of *Poisson* random variables.

The Poisson distribution: basic framework

More precisely, if:

- The basic events occur one at a time (never simultaneously);
- The occurrence of a basic event in a given period is independent of the occcurence of the event in any other non-overlapping period;
- The expected # of basic events during any period of unit length is μ, and the expected # of basic events during any period of length t is $t\mu$;

then $X =$ # occurrences of the basic event in a period of unit length is a Poisson random variable with parameter (and mean) μ.

The Poisson distribution: the pdf

The pdf of a Poisson random variable with parameter μ is given by

$$f(x) = \frac{e^{-\mu}\mu^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots,.$$

Example: Suppose that X, the number of hurricanes in any given calendar year, is a Poisson random variable with mean 5. Find the pdf of X and use it to determine the probability that there will be 3 or fewer hurricanes in 2016.

Answer:

$$\begin{aligned} f(x) &= \frac{e^{-5}5^x}{x!}, \quad \text{for } x = 0, 1, 2, \ldots,, \\ \text{so } P(X \leq 3) &= e^{-5}\left(\frac{5^0}{0!} + \frac{5^1}{1!} + \frac{5^2}{2!} + \frac{5^3}{3!}\right) = .2650. \end{aligned}$$

The Poisson distribution: the CDF and Table C.2

The CDF of a Poisson random variable with parameter μ is given by

$$F(d) = P(X \leq d) = \sum_{x=0}^{d} \frac{e^{-\mu}\mu^{x}}{x!}, \quad \text{for } d = 0, 1, 2, \ldots,.$$

Hurricane example, continued: Use Table C.2 to re-obtain $P(X \leq 3)$ and also to obtain the probability that there are exactly 3 hurricanes in 2016 and the probability that there is at least one hurricane in 2016.

The Poisson distribution: changing the length of the time period

The third part of the framework for a Poisson random variable (p. 98 of these notes) tells us that if the number of basic events in a period of unit length has expected value μ, then the number of basic events in a period of length t is a Poisson random variable with parameter (and mean) $t\mu$. So, with proper modification we can compute probabilities for events involving a period of any length.

Hurricane example, continued: What is the probability that there are 10 or fewer hurricanes from 2016-2018 (inclusive)?

Poisson approximation to the binomial distribution

In addition to being useful for determining probabilities involving Poisson random variables, the Poisson pdf (and CDF) is useful for approximating probabilities involving binomial random variables in certain circumstances. The circumstances are:

$$n \geq 100, \qquad np \leq 10.$$

In place of the binomial CDF with parameters n and p, we use the Poisson CDF with mean $\mu = np$.

Poisson approximation to the binomial distribution

Example: Suppose that you go on a trip to the Caribbean, and while you are there, each time you are bitten by a mosquito the probability that the mosquito is carrying the Zika virus is .01. Suppose further that you are bitten by 120 mosquitos while on the trip. What is the probability that at least one of these bites will be from a Zika virus carrier?

Continuous random variables

Recall that if X is a continuous random variable, it can take on any value in a specified interval. Higher mathematics tells us that the # of real numbers in an interval is infinite, in fact *uncountably infinite* (in contrast to, say, the nonnegative integers which are *countably infinite*). As a result, the axioms of probability dictate that:

- We must assign a probability of 0 to the event that X equals any single real number in the specified interval;
- The only events that can be assigned meaningful nonzero probabilities are subintervals (or unions thereof) of the specified interval.

Continuous random variables

Thus, if a and b are two real numbers such that $a < b$, then

$$P(X = a) = 0 \qquad \text{and} \qquad P(X = b) = 0,$$

but $P(a < X < b)$ can be nonzero. Furthermore,

$$P(a < X < b) = P(a \leq X < b) = P(a < X \leq b) = P(a \leq X \leq b)$$

since, for example,

$$P(a \leq X < b) = P(X = a) + P(a < X < b) = 0 + P(a < X < b).$$

Example: If we randomly select an incoming UI freshman, we must assign 0 to the probability that his/her HS GPA equals 2.93, and ...

Continuous random variables: the pdf

Let X be a continuous random variable. Since $P(X = x) = 0$ for any number x, the pdf of X must be defined somewhat differently than it was for a discrete random variable.

The pdf of a continuous random variable is a function, f, that satisfies:

- $f(x) \geq 0$;
- the area under the graph of f (and above the x-axis) is equal to 1.0;
- for any real numbers a and b with $a \leq b$, $P(a \leq X \leq b)$ is given by the area under the graph of f between $x = a$ and $x = b$.

Continuous random variables: the pdf

Example 1: Suppose that X is a continuous random variable with pdf given by the following picture:

Equivalently, $\quad f(x) = \begin{cases} \frac{1}{2}, & \text{if } 1 < x < 3 \\ 0, & \text{otherwise.} \end{cases}$

Then, e.g., $P(1 < X < 2) = \frac{1}{2} \cdot 1 = \frac{1}{2}$, $P(1.5 < X < 4) =$

We can also solve problems like: Find a number c such that $P(X < c) = \frac{2}{3}$.

Continuous random variables: the pdf

Example 2: Suppose that X is a continuous random variable with pdf given by the following picture:

Then, e.g.:

- $P(X > 1) =$
- $P(|X| < 1) =$
- Find a number c such that $P(0 < X < c) = \frac{1}{4}$:

Continuous random variables: the CDF

If X is a continuous random variable, its CDF is given by

$$F(x) = P(X \leq x),$$

which is the area under the graph of the pdf from $-\infty$ to x.

For the previous Example 1,

$$F(x) = \begin{cases} 0 & \text{for } x < 1, \\ \frac{1}{2}(x-1) & \text{for } 1 \leq x < 3, \\ 1 & \text{for } x \geq 3. \end{cases}$$

Continuous random variables: mean and variance

The mean, μ, of the distribution of a continuous random variable is the place where a fulcrum placed under the graph of the pdf would make the graph balance.

In the previous Example 1, $\mu = 2$; in the previous Example 2, $\mu = 0$. These were easily determined because of the symmetry of the pdf around its mean. More generally, the mean is computed by methods of integral calculus (never mind!).

The variance, σ^2, of the distribution of a continuous random variable is a measure of how spread out the pdf is, and it's also computed by methods of integral calculus (never mind!).

The normal distribution: Introduction

The most important continuous probability distribution is the **normal distribution**, whose pdf is a bell-shaped curve. Why is it so important?

- In practice, samples of many physical measurements and other biological variables have a relative frequency distribution which often seems to be bell-shaped.
- The normal distribution has nice mathematical properties.
- The normal distribution provides an accurate approximation to the distribution of the sample mean, no matter what the distribution of the observations in the sample is (Central Limit Theorem — much more on this later).

The normal distribution: the pdf

The normal distribution's pdf, the "normal curve," has the following features:

- a single peak, located at the mean μ
- symmetric around μ (thus mean=median=mode)
- tails that extend infinitely far in both directions
- a standard deviation, σ, that controls where the curve's 2 points of inflection are: $\mu + \sigma$ and $\mu - \sigma$
- a complicated functional form (given in text — never mind!)

The normal distribution: the pdf

There is a different normal curve for each set of parameters μ and σ^2, which we label as $\mathrm{N}(\mu, \sigma^2)$.

Example 1: $\mathrm{N}(0,1)$, $\mathrm{N}(1,1)$, $\mathrm{N}(2,1)$

Example 2: $\mathrm{N}(0,1)$, $\mathrm{N}(0,4)$, $\mathrm{N}(0,9)$

The standard normal distribution and its CDF

The specific normal curve, $\mathrm{N}(0,1)$, is called the *standard* normal distribution, and the corresponding random variable is denoted by Z. We write

$$Z \sim \mathrm{N}(0,1).$$

The CDF of the standard normal distribution is

$$F(z) = P(Z \leq z),$$

and is given by the area under the standard normal curve to the left of z. These probabilities are listed in Table C.3 for a large number of values of z.

The standard normal distribution: Determining probabilities

We can use Table C.3 and knowledge of the symmetry of the standard normal curve around 0 to obtain the probabilities of many events involving Z.

- $P(Z < 0) =$
- $P(Z > 0) =$
- $P(Z < -1.76) =$
- $P(Z > -1.76) =$
- $P(Z < 0.62) =$

The standard normal distribution: Determining probabilities

- $P(-0.39 < Z < 1.64) =$
- $P(0.50 < Z < 1.50) =$
- $P(|Z| < 1.00) =$
- $P(|Z| < 2.00) =$
- $P(|Z| < 3.00) =$

Drawing a picture is always a good idea.

The standard normal distribution: Inverse problems

We can also solve "inverse" problems, where we are given the probability and asked to find a z-value. For example, find z such that:

- $P(Z < z) = 0.9750$
- $P(Z > z) = 0.7486$
- $P(0 < Z < z) = 0.3749$
- $P(-z < Z < z) = 0.9010$

The normal distribution: Standardization

In practice, biological variables rarely have a standard normal distribution, but may instead have some other normal distribution. How do we determine probabilities of events of interest in this case?

Example: Suppose that diastolic blood pressures of hypertensive women are normally distributed, with mean 100 mm Hg and standard deviation 16 mm Hg. What is the probability that the diastolic blood pressure of a randomly selected hypertensive woman is less than 90 mm Hg?

We want $P(X < 90)$ where $X \sim \mathrm{N}(100, 16^2)$. We use *standardization* to convert this type of problem to an equivalent one involving Z.

The normal distribution: Standardization

Standardization requires subtracting μ from X and dividing the result by σ, and when we do so, the resulting random variable has a standard normal distribution, i.e.

$$\frac{X-\mu}{\sigma} \sim \mathrm{N}(0,1).$$

Applied to the previous example, we have

$$P(X<90) = P\left(\frac{X-100}{16} < \frac{90-100}{16}\right) = P(Z<-0.63) = 0.2643.$$

The normal distribution: Standardization

We can also do "inverse" problems like the following: 10% of the population of hypertensive women have a diastolic blood pressure above what level?

We want a number (in units mm Hg), say c, such that $P(X > c) = 0.10$.

Normal approximation to the binomial distribution

In addition to being useful for determining probabilities involving normal random variables, the normal CDF is useful for approximating probabilities involving binomial random variables in certain circumstances. The circumstances are:

$$np > 5, \qquad n(1-p) > 5.$$

In place of the binomial CDF with parameters n and p, we use the normal CDF with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

Normal approximation to the binomial distribution

Also, since we are approximating a discrete distribution with a continuous one, we often employ a *continuity correction factor*. For a random variable whose possible values are all integers within some interval, this involves subtracting 0.5 from x in events of the form $(X < x)$, and adding 0.5 to x in all events of the form $(X \leq x)$. So instead of finding $P(X < x)$ we find $P(X < x - 0.5)$, and instead of finding $P(X \leq x)$ we find $P(X < x + 0.5)$.

Normal approximation to the binomial distribution

Example: Recall the Caribbean trip you take, on which you are bitten by 120 mosquitos. Suppose that each time you are bitten by a mosquito on the trip, the probability that the mosquito is carrying the Zika virus is .10. What is the probability that at least one of the 120 bites will be from a Zika virus carrier?

Some practice combining concepts of Chapters 2 & 3

Example 1: The proportion of U.S. residents that are lactose intolerant is believed to be about 0.10. If this is correct and we randomly sample 7 people from the U.S. population, (a) what is the probability that fewer than 3 and an odd number of people in the sample are lactose intolerant?

(b) What is the probability that fewer than 3 or an odd number of people in the sample are lactose intolerant?

(c) What is the probability that fewer than 3 people in the sample are lactose intolerant, given that an odd number of people in the sample are lactose intolerant?

(d) What is the probability that an odd number of people in the sample are lactose intolerant, given than fewer than 3 people in the sample are lactose intolerant?

Example 2: Suppose $Z \sim \mathrm{N}(0,1)$. (a) Find $P(0.50 < Z < 1.50 \cap Z < 1.00)$.

(b) Find $P(0.50 < Z < 1.50 | Z < 1.00)$.

(c) In a random sample of size 6 from a population whose distribution is $\mathrm{N}(0, 1)$, what is the probability distribution of X, where X is defined as the number of observations in the sample that are less than 1.00?

Sampling distributions: Introduction

Now we begin to wed the concepts of sample statistics (Chapter 1) and probability distributions (Chapters 2 and 3).

Recall that we take a (random) sample from a population of interest because we want to estimate some parameter of that population; we use the sample-based estimate (e.g. $\overline{X}$) to make inferences about the population parameter (e.g. μ). We want to be able to say something about how close our estimate is to the parameter (e.g. $P(|\overline{X} - \mu| < .01) = ?$).

To do so, we need to know the *sampling distribution*, i.e. the probability distribution of the sample-based estimate.

Sampling distributions: Introduction

Suppose the population of interest is newborn infants in Iowa; the random variable of interest is their weight at birth, X; and the parameter of interest is their mean, μ. Suppose we propose to take a random sample $X_1, X_2, \ldots, X_n$ (of size n) from the population and compute the sample mean, $\overline{X}$, as an estimate of μ. What are the random experiment and random variable here?

If you took a random sample of size n and I did likewise, would we be likely to obtain the same value of $\overline{X}$?

Thus $\overline{X}$ is a random variable in its own right, and it has a probability distribution. What is this probability distribution?

Sampling distribution of $\overline{X}$: A discrete example

Suppose X is a discrete random variable with the following pdf:

x	$f(x)$
1	.4
2	.1
3	.2
4	.3

$$\mu_X = 2.40, \quad \sigma_X^2 = 1.64$$

Now think of the x's as the values of objects in some very large population, and the $f(x)$'s as their corresponding relative frequencies. Imagine taking a random sample of size 2 from this population with replacement, and let $\overline{X}$ represent the mean of this sample. Then the probability distribution of $\overline{X}$ is as follows:

Sampling distribution of $\overline{X}$: A discrete example

$\overline{x}$	$f(\overline{x})$
1.0	(.4)(.4)=.16
1.5	(.4)(.1)+(.1)(.4)=.08
2.0	(.4)(.2)+(.2)(.4)+(.1)(.1)=.17
2.5	(.4)(.3)+(.3)(.4)+(.1)(.2)+(.2)(.1)=.28
3.0	(.1)(.3)+(.3)(.1)+(.2)(.2)=.10
3.5	(.2)(.3)+(.3)(.2)=.12
4.0	(.3)(.3)=.09

$$\mu_{\overline{X}} = 2.40, \quad \sigma^2_{\overline{X}} = 0.82$$

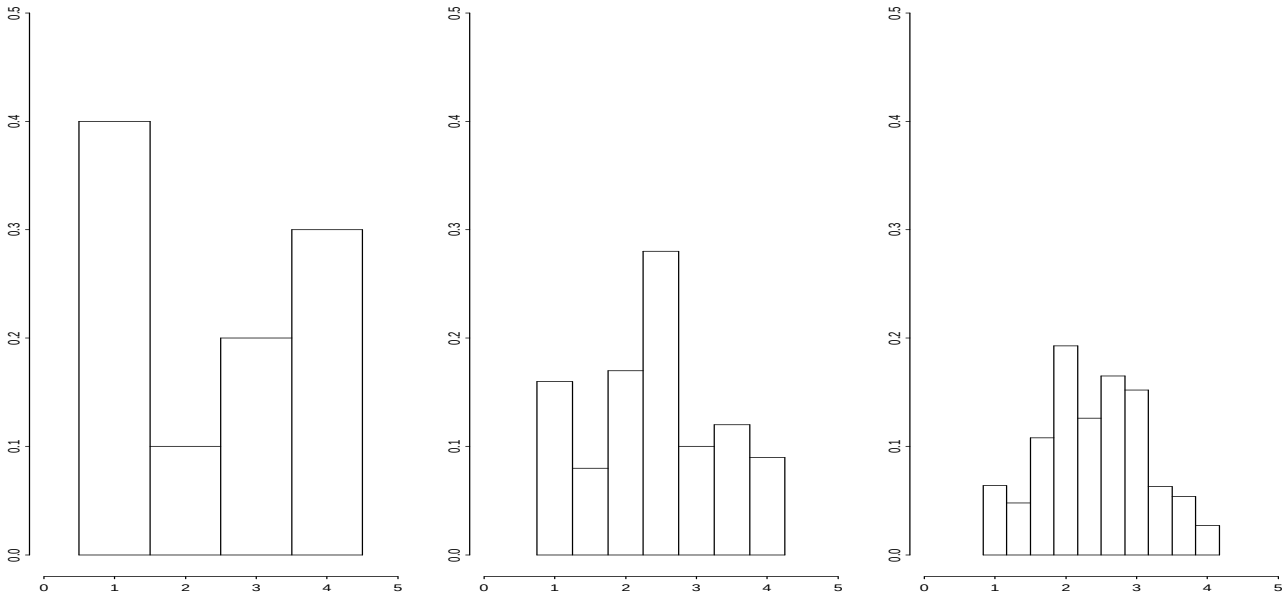
Sampling distribution of $\overline{X}$: A discrete example

If we take a random sample of size 3 (rather than 2), then the pdf of $\overline{X}$ is:

$\overline{x}$	$f(\overline{x})$
1.0	.064
$1.\overline{3}$	.048
$1.\overline{6}$	.108
2.0	.193
$2.\overline{3}$	.126
$2.\overline{6}$	.165
3.0	.152
$3.\overline{3}$	.063
$3.\overline{6}$	.054
4.0	.027

$\mu_{\overline{X}} = 2.40, \quad \sigma^2_{\overline{X}} = 0.54\overline{6}$

Sampling distribution of $\overline{X}$: A discrete example



Sampling distribution of $\overline{X}$: A discrete example

If we take a random sample of size 100, then the pdf of $\overline{X}$ is:

$\mu_{\overline{X}} = 2.40, \quad \sigma^2_{\overline{X}} = 0.0164$, and the pdf is very similar to that of a normal distribution.

Sampling distribution of $\overline{X}$: A discrete example

As the sample size increases, how does the distribution of $\overline{X}$ behave?

- It has the same mean as the distribution of X, regardless of n.
- Its variance is smaller than the variance of X, and keeps getting progressively smaller.
- It becomes progressively more bell-shaped (more normal).

The behavior of the sampling distribution of $\overline{X}$ seen in the particular example above also occurs much more generally. The general result is known as the **Central Limit Theorem**.

The Central Limit Theorem (CLT)

When taking a random sample of size n from a population with any probability distribution having mean μ_X and variance σ_X^2, the distribution of $\overline{X}$:

1. has mean μ_X;
2. has variance σ_X^2/n;
3. becomes more and more like a normal distribution as n increases. (If the population you're sampling from is normal, then the distribution of $\overline{X}$ is exactly normal.)

Amazing!

The Central Limit Theorem (CLT)

We can summarize the CLT as follows:

$$\overline{X} \dot{\sim} \mathrm{N}(\mu, \frac{\sigma^2}{n}) \quad \text{for } n \text{ sufficiently large.}$$

Or equivalently,

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \dot{\sim} \mathrm{N}(0,1) \quad \text{for } n \text{ sufficiently large.}$$

The quantity $\sigma/\sqrt{n}$ is called the *standard error of the mean*.
Note: as the sample size increases, the standard error of the mean decreases.

The Central Limit Theorem in practice

As a consequence of the CLT, we can approximate probabilities of events involving $\overline{X}$ using a normal distribution, provided the sample size (n) is sufficiently large.

How large must n be to safely use the approximation?

- If the sampled population is normal, then any n is OK.
- If the sampled population is non-normal but symmetric, then $n \geq 5$ suffices.
- If the sampled population is not symmetric, then $n \geq 25$ suffices in all but some pathological cases that usually don't occur in practice.

The CLT in practice: Examples

Suppose that the birth weights of Iowa infants born at gestational age 40 weeks are approximately normally distributed with mean $\mu = 3500$ g and standard deviation $\sigma = 430$ g.

- What is the (approximate) probability that the birth weight of an infant randomly selected from this population is less than 3000 g?

The CLT in practice: Examples

- What is the (approximate) probability that the average birth weight of 5 infants randomly selected from this population is less than 3000g?

The CLT in practice: Examples

- What birth weight (approximately) cuts off the lower 5% of the distribution of this population's birth weights?

The CLT in practice: Examples

- What birth weight (approximately) cuts off the lower 5% of the distribution of sample mean birth weights based on samples of size 5 drawn from this population?

Interval estimation: Introduction

A *point estimate* of a population parameter is a single number, computed from a sample, believed to be close to the parameter. Examples:

- $\overline{X}$ is a point estimate of μ
- s^2 is a point estimate of σ^2

An *interval estimate*, or *confidence interval*, is an interval of numbers that is computed from a sample in such a way that the interval has a prespecified probability of containing the population parameter.

Confidence interval for μ: Derivation

Suppose a random sample is taken from a population with unknown mean μ and known variance σ^2, and n is large enough for the CLT to apply. Then, approximately,

$$P(-2.58 \leq \frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \leq 2.58) = 0.99.$$

(The interval from -2.58 to +2.58 captures the middle 99% of the standard normal distribution.)

Algebraic manipulations to get μ by itself:

Confidence interval for μ: Derivation

So the probability that the random interval

$$(\overline{X} - 2.58\frac{\sigma}{\sqrt{n}}, \overline{X} + 2.58\frac{\sigma}{\sqrt{n}})$$

contains μ is 0.99. When we substitute the observed sample mean for $\overline{X}$, we call the interval a 99% confidence interval for μ and we say we are 99% confident that μ lies within this interval.

Note that the width of the interval is not random: it =

Confidence interval for μ: Example

Ten insomniac patients were given two sleep-inducing drugs, say drugs A and B, on separate occasions when they were having trouble sleeping. The additional hours of sleep gained by using drug B instead of drug A, for the ten patients, are given below (in hours):

$$1.2, 2.4, 1.3, 1.3, -1.0, 3.8, 0.0, 0.8, 4.6, 1.4$$

For these data, $\overline{X} = 1.58$ hrs. Assume that $\sigma = 1.5$ hrs and that the distribution of sleep gain (B over A) is symmetric. Find an approximate 99% confidence interval for μ, the mean additional hours of sleep gained by using drug B rather than drug A among all insomniacs.

Answer:

Confidence interval for μ: Levels of confidence

There is nothing special about 99% as a level of confidence; if we wish, we can obtain a confidence interval for μ with greater or lesser level of confidence.

In general, if $0 < \alpha < 1$, a $100(1-\alpha)\%$ confidence interval for μ is given by

$$(\overline{X} - z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}})$$

where $z_{1-\alpha/2}$ cuts off the upper $(\alpha/2)\%$ of the $\mathrm{N}(0,1)$ distribution.

Confidence level	α	$z_{1-\alpha/2}$
90%	.10	1.645
95%	.05	1.96
99%	.01	2.58

Confidence interval for μ: Levels of confidence

Insomniac example: A 95% CI for μ is

Is this interval narrower or wider than the 99% CI computed on page 150 of these notes?

Confidence interval for μ: Unknown σ

Up to now, we've assumed (unrealistically!) that we know σ. In practice, we generally don't. So how to obtain a CI when we don't know σ?

A natural idea: replace σ in the CI formula with s, its point estimate, obtained from the same sample used to obtain the point estimate $\overline{X}$ of μ:

$$(\overline{X} - z_{1-\alpha/2}\frac{s}{\sqrt{n}}, \overline{X} + z_{1-\alpha/2}\frac{s}{\sqrt{n}})$$

While this is not completely off-base, it is not quite right because the distribution of $\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}}$ is not standard normal. Instead, it has what is called a *t distribution with* $n-1$ *degrees of freedom*.

Diversion: The t distributions

- Like the $\mathrm{N}(0,1)$, the t distributions are continuous, symmetric, bell-shaped, with mean 0.
- A different t distribution exists for each sample size, n, or equivalently for each degrees of freedom, $n-1$.
- The t distribution with any finite degrees of freedom is more spread out than the $\mathrm{N}(0,1)$.
- The larger that n (or $n-1$) is, the more closely the t distribution resembles the $\mathrm{N}(0,1)$.
- The t distributions are tabled in Table C.4.

Confidence interval for μ: Unknown σ

So the proper expression for a $100(1-\alpha)\%$ CI for μ when σ is unknown is

$$(\overline{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}, \overline{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}})$$

where $t_{1-\alpha/2,n-1}$ is a number (from Table C.4) that cuts off the upper $(\alpha/2)\%$ of the probability of the t distribution with $n-1$ degrees of freedom.

This is an exact $100(1-\alpha)\%$ CI when the random sample is drawn from a normally-distributed population. If the population is not normally distributed, then the interval given above is an adequate approximate $100(1-\alpha)\%$ CI provided that the sample size is sufficiently large (see p. 142 for how large is "large enough").

Confidence interval for μ: Example with unknown σ

The insomniac example, revisited: Ten insomniac patients were given two sleep-inducing drugs, say drugs A and B, on separate occasions when they were having trouble sleeping. The additional hours of sleep gained by using drug B instead of drug A, for the ten patients, are given below (in hours):

$$1.2, 2.4, 1.3, 1.3, -1.0, 3.8, 0.0, 0.8, 4.6, 1.4$$

For these data, $\overline{X} = 1.58$ hrs and $s = 1.66$ hrs. Assume that the distribution of sleep gain (B over A) is symmetric. Find an approximate 99% confidence interval for μ, the mean additional hours of sleep gained by using drug B rather than drug A among all insomniacs.

Answer:

Confidence interval for μ: Factors affecting width

The narrower the confidence interval, the more precisely we've pinned down μ. The width of a $100(1-\alpha)\%$ confidence interval for μ is

What factors affect width?

- Level of confidence
- Sample standard deviation
- Sample size

Confidence interval for σ^2: Introduction

Now suppose we are interested in an interval of plausible values for not (or not only) the population mean, but for the population variance, σ^2. We've already introduced a point estimate for σ^2, namely the sample variance s^2.

By analogy with the CI for the mean, we need to find a mathematical expression for a random variable that contains σ^2 and has a known probability distribution. In theoretical statistics courses, it is proved that if we take a random sample of size n from a normal distribution, then the quantity

$$\frac{(n-1)s^2}{\sigma^2}$$

has a known probability distribution called the chi-square distribution (with $n-1$ degrees of freedom).

Diversion: The chi-square (χ^2) distribution

- a probability distribution for a continuous random variable
- the pdf is positive over the interval $(0, \infty)$
- the pdf is asymmetrically bell-shaped (right-skewed)
- there is a different chi-square distribution for each value of a parameter called the degrees of freedom (like the t distribution in this respect). We label each such distribution χ^2_{df}.
- The CDF of χ^2_{df} is tabled in Table C.5 for $df = 1, \ldots, 60$.

Confidence interval for σ^2: Derivation

Write $\chi^2_{n-1,\alpha/2}$ and $\chi^2_{n-1,1-\alpha/2}$ for the values that cut off $100\alpha/2\%$ from each tail of the χ^2_{n-1} distribution (leaving $100(1-\alpha)\%$ of the distribution in the middle). Then

$$P\left(\chi^2_{n-1,\alpha/2} \leq \frac{(n-1)s^2}{\sigma^2} \leq \chi^2_{n-1,1-\alpha/2}\right) = 1-\alpha$$

which can be manipulated algebraically (see textbook) to yield

$$P\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}} \leq \sigma^2 \leq \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}}\right) = 1-\alpha.$$

Confidence interval for σ^2: Formula

Thus, a $100(1-\alpha)\%$ confidence interval for σ^2, assuming the randomly sampled population is normally distributed, is given by

$$\left(\frac{(n-1)s^2}{\chi^2_{n-1,1-\alpha/2}}, \frac{(n-1)s^2}{\chi^2_{n-1,\alpha/2}} \right).$$

Observe that this interval is <u>not</u> of the form "$s^2 \pm$ something". So s^2 is not the midpoint of the CI.

Confidence interval for σ^2: Example

The lifespans of a population of men having a certain gene are normally distributed. A random sample of size 16 from this population had an average lifespan of 81.2 years with a standard deviation of 8.0 years. Determine a 95% confidence interval for the variance of lifespans for this population of men.

Confidence interval for a proportion: Introduction

Now suppose the variable of interest for the population of interest is categorical, and even more specifically, dichotomous; that is, the values of the variable are coded as either a 1 (for "success," i.e. possessing the characteristic of interest) or 0 (for "failure," i.e. not possessing the characteristic of interest). In this case the population parameter of interest is p, the proportion of the population having the characteristic of interest.

From a random sample $X_1, X_2, \ldots, X_n$ drawn from the population, we estimate p by the sample proportion of successes,

$$\hat{p} = \frac{\#\text{ successes}}{n} = \frac{\sum_{i=1}^{n} X_i}{n} = \overline{X}.$$

Confidence interval for a proportion: Sampling distribution of $\hat{p}$

Since $\hat{p}$ is a sample mean, the CLT gives its approximate sampling distribution. Furthermore, recall that the expressions for the mean and variance of a binomial random variable (with parameters n and p) are np and $np(1-p)$, respectively. Thus the mean and variance of $\hat{p}$ are p and $p(1-p)/n$, respectively. Thus by the CLT,

$$\hat{p} \dot{\sim} \mathrm{N}\left(p, \frac{p(1-p)}{n}\right) \quad \text{for } n \text{ sufficiently large.}$$

Here, "sufficiently large" means $np \geq 5$ and $n(1-p) \geq 5$.

Standardizing yields

$$\frac{\hat{p}-p}{\sqrt{p(1-p)/n}} \dot{\sim} \mathrm{N}(0,1).$$

Confidence interval for a proportion

Using the result at the bottom of the previous page, we can derive the following approximate $100(1-\alpha)\%$ CI for p:

$$(\hat{p} - z_{1-\alpha/2}\sqrt{p(1-p)/n},\ \hat{p} + z_{1-\alpha/2}\sqrt{p(1-p)/n}).$$

But notice that the standard error of $\hat{p}$ depends on p, which is unknown! To fix this, we simply replace p with $\hat{p}$ in the expression for the standard error of $\hat{p}$, yielding the following approximate $100(1-\alpha)\%$ CI for p:

$$(\hat{p} - z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n},\ \hat{p} + z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}).$$

Confidence interval for a proportion: Example

In 2001-02, 272 deer were legally killed by hunters in the Mount Horeb area of SW Wisconsin. From tissue sample analysis, it was determined that 9 of the deer had chronic wasting disease (a disease similar to mad cow disease). Determine a 95% confidence interval for the proportion of the entire deer population in this area of Wisconsin that had chronic wasting disease in 2001-02.

Answer:

Any issues with the assumptions?

Confidence interval for a proportion: Sample size considerations

The width of the $100(1-\alpha)\%$ CI for p is

$$2z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$$

and the *margin of error*, defined as half the width of this CI, is

$$z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}.$$

We can use these expressions (with some variations) to determine the sample size necessary for the width or margin of error of this CI to be less than a pre-specified value.

Confidence interval for a proportion: Sample size considerations

Example: Suppose we wish to obtain a 95% CI for the proportion of lefthanded people in the U.S., and we want this CI to be of width .02 or less (equivalently, we want the CI's margin of error to be .01 or less). Using the formula above, we want to choose n such that

$$2(1.96)\sqrt{\hat{p}(1-\hat{p})/n} \leq .02,$$

or equivalently,

$$n \geq \frac{4(1.96)^2\hat{p}(1-\hat{p})}{.02^2}.$$

Confidence interval for a proportion: Sample size considerations

But what do we use for $\hat{p}$? Three possibilities:

- *Prior guess*. If we guess that the proportion of lefthanders is about 10%, then we have

$$n \geq \frac{4(1.96)^2(0.1)(0.9)}{.02^2} = 3457.44.$$

- *Pilot study*. Suppose we sample 200 people to start with, and 32 of them are lefthanded. Then our provisional $\hat{p}$ is $32/200 = 0.16$; using this we obtain

$$n \geq \frac{4(1.96)^2(0.16)(0.84)}{.02^2} = 5163.11.$$

Confidence interval for a proportion: Sample size considerations

- *Conservative approach*. Note that the function $f(x) = x(1-x)$ is maximized over $0 < x < 1$ at $x = 0.5$; so replace $\hat{p}$ with 0.5:

$$n \geq \frac{4(1.96)^2(0.5)(0.5)}{.02^2} = 9604.$$

Interpretations of confidence intervals

There are correct interpretations of confidence intervals, and there are incorrect interpretations. Consider a $100(1-\alpha)\%$ confidence interval for a population mean μ, based on a random sample of size n. (Similar comments apply to a confidence interval for p.) Suppose the computed interval is (117.3,126.9).

Correct interpretations:

- If we were to repeatedly take random samples of size n from the population, on average $100(1-\alpha)\%$ of the CI's so constructed would contain μ.
- We are $100(1-\alpha)\%$ confident that (117.3,126.9) contains μ.

Interpretations of confidence intervals

Incorrect interpretations:

- The probability that (117.3,126.9) contains μ is $1-\alpha$.
- $100(1-\alpha)\%$ of the population's values lie in (117.3,126.9).
- $100(1-\alpha)\%$ of the sample means (based on samples of size n) lie in (117.3,126.9).

Hypothesis testing: Introduction

Confidence intervals are one of the two main currencies of statistical inference. The other is *hypothesis testing*.

Briefly, statistical hypothesis testing is a procedure for testing the validity of a claim about a population parameter by evaluating how compatible, probabilistically speaking, it is with a relevant statistic computed from a random sample.

Example: An investigator doesn't know the population mean body temperature of African elephants, but using current knowledge of the physiology, surface area, and weight of African elephants, together with current theory of how these things affect body temperature, he hypothesizes that it is 96.0°F. If the mean of a random sample of 50 African elephants is 96.2°F, does this cast doubt that the current theory is applicable to African elephants?

Hypothesis testing: Null and alternative hypotheses

We've answered a few questions similar to this in an informal way (e.g. Problems 1.4b, 4.23b). Now, however, we formalize things so that we can test hypotheses in a consistent, scientifically valid way.

We consider only those situations in which there are two mutually exclusive and exhaustive hypotheses:

1. *Null hypothesis*, H_0. This is the default, or status quo, claim about a population parameter.

2. *Alternative hypothesis*, H_a. Also called the *research hypothesis*, this is the scientific investigator's claim about the population parameter.

Hypothesis testing: Null and alternative hypotheses

- The investigator must specify these two hypotheses according to the problem at hand and his/her goals.
- Depending on their goals, one investigator's H_0 may be different than another investigator's H_0, even for the same problem.
- The burden of proof is always on the investigator to provide strong evidence that the null hypothesis is false (this is consistent with the scientific method).

Example 1 (not biological): A murder is committed, and a suspect is arrested and put to trial.

The jury's H_0:

The jury's H_a:

Hypothesis testing: Null and alternative hypotheses

Example 2: Mean body temperature of African elephants. Let μ represent the mean body temperature for the population of elephants.

The researcher's H_0:

The researcher's H_a:

Example 3: A new treatment regimen for pancreatic cancer is developed by researchers. Let p represent the proportion of a conceptual population of patients receiving the new treatment that will be alive after 5 years, and suppose that this proportion for the population of patients receiving the current best treatment is 0.046.
The researcher's H_0:

The researcher's H_a:

Hypothesis testing: Hypotheses about a population mean

Hypotheses about a population mean are of 3 types:

- H_0: $\mu = \mu_0$ versus $H_a : \mu \neq \mu_0$
- H_0: $\mu \leq \mu_0$ versus $H_a : \mu > \mu_0$
- H_0: $\mu \geq \mu_0$ versus $H_a : \mu < \mu_0$

In the first of these, H_a is *two-sided*; in the others, H_a is *one-sided*.

Note that μ_0 is always included in the null hypothesis.

Hypothesis testing: Type I and Type II errors

Based on the evidence, H_0 will be either rejected or not rejected (accepted). This decision can be right or wrong. There are two types of correct decisions, and two types of wrong decisions (errors); the latter are called Type I and Type II errors.

Example 1: Jury's decision about murder suspect

Hypothesis testing: Type I and Type II errors

Example 2: Researcher's decision about population mean body temperature of African elephants

Hypothesis testing: Type I and Type II errors

Define:

$$\begin{aligned} \alpha &= \text{Probability of making a Type I error,} \\ \beta &= \text{Probability of making a Type II error.} \end{aligned}$$

- In statistical hypothesis testing, we have the ability to set α at some relatively small number, traditionally at .05 or .01; then we take what we get for β (generally, the smaller we make α, the larger that β gets).
- Some situations may call for a larger or smaller α, e.g. desperate health situations.
- α is also called the *level of significance*.

Hypothesis testing: Power

The *power* of a test is the probability, using that test, that we will reject H_0 when it is false.

- Thus, Power $= 1 - \beta$.
- High power is a good thing!
- Some tests have higher power than others, but often at the price of more restrictive assumptions

Hypothesis testing: Six-step procedure

1. Formulate H_0 and H_a, based on the scientific question of interest.

2. Choose a level of significance, α, based on the relative importance of Type I and Type II errors in the given situation.

3. Choose an appropriate *test statistic* for the problem and compute it. We generally choose the most powerful test statistic available, provided that the assumptions required for its validity are satisfied.

4. (a) Determine a *critical value(s)*, using a table, to which the test statistic's value will be compared; OR

   (b) Determine the *P value*, using a table, to compare to α.

Hypothesis testing: Six-step procedure

5. (a) If the test statistic is more extreme than the critical value(s), then reject H_0; otherwise do not reject H_0. OR

   (b) If the P value is less than α, reject H_0; otherwise do not reject H_0.

6. Express your conclusion as an answer to the scientific question of interest.

Test statistics for hypotheses about a population mean

A test statistic is a quantity, computable from a random sample, which measures the discrepancy between what the data say about the population parameter's value and what H_0 claims the population parameter's value is.

For testing hypotheses about a population mean μ, this discrepancy can (initially) be measured by

$$\overline{X} - \mu_0.$$

However, the "extremeness" of any particular value of this discrepancy cannot be judged until it is scaled by the inherent variability of the data. So for our test statistic we scale this discrepancy by the variability of $\overline{X}$, or an estimate thereof.

Test statistics for hypotheses about a population mean

According to the CLT, the standard error of $\overline{X}$ is $\sigma/\sqrt{n}$. So, if σ is known to us, we may use as our test statistic the "z-statistic"

$$Z = \frac{\overline{X} - \mu_0}{\sigma/\sqrt{n}}.$$

- Dividing by $\sigma/\sqrt{n}$ calibrates the discrepancy between $\overline{X}$ and μ_0 in units of standard error.
- Note that if μ really does equal μ_0 (i.e. if H_0 is true), and if n is sufficiently large, then by the CLT the distribution of the z-statistic is approximately $\mathrm{N}(0,1)$.

Test statistics for hypotheses about a population mean

If σ is not known (which is almost always the case in practice), then we may replace it with s in the z-statistic, yielding as a test statistic the "t-statistic"

$$t = \frac{\overline{X} - \mu_0}{s/\sqrt{n}}.$$

- The discrepancy between $\overline{X}$ and μ_0 is calibrated in units of estimated standard error.
- If μ really does equal μ_0 (i.e. if H_0 is true), and if n is sufficiently large, then the distribution of the t-statistic is approximately t with $n-1$ degrees of freedom.

Critical values for testing hypotheses about a population mean

On the previous two pages, we noted what the distributions were, under H_0, for the z and t test statistics. Tables of these distributions are where we go to find critical values.

For the z-statistic, critical values are:

- $\pm z_{1-\alpha/2}$ if H_a is two-sided
- $z_{1-\alpha}$ if H_a is $\mu > \mu_0$
- z_α if H_a is $\mu < \mu_0$

Critical values for testing hypotheses about a population mean

For the t-statistic, critical values are:

- $\pm t_{1-\alpha/2,n-1}$ if H_a is two-sided
- $t_{1-\alpha,n-1}$ if H_a is $\mu > \mu_0$
- $t_{\alpha,n-1}$ if H_a is $\mu < \mu_0$

If our computed test statistic is more extreme than the critical value, we reject H_0; otherwise, we do not reject H_0.

Hypothesis testing for a population mean: Example

A researcher wanted to test the hypothesis that the mean body temperature of African elephants was 96.0°F. He has no prior notion about which direction the mean body temperature of elephants will differ from 96.0°F if it is not equal to 96.0.

- *Step 1.* So, letting μ represent the mean body temperature of African elephants, he wants to test

$$H_0 : \mu = 96.0 \quad \text{versus} \quad H_a : \mu \neq 96.0.$$

- *Step 2.* He makes a traditional choice of $\alpha = .05$.

A random sample of 50 African elephants is taken, from which the sample mean and sample standard deviation were computed as follows: $\overline{X} = 96.2$, $s = 0.63$.

Hypothesis testing for a population mean: Example

- *Step 3.* Since the population standard deviation is not known, but the sample size is sufficiently large, he chooses the t-statistic as our test statistic and computes it as follows:

- *Step 4.* Critical values are $\pm t_{.975,49} = \pm 2.010$.
- *Step 5.* Since the computed test statistic is more extreme than the critical values, he rejects H_0.
- *Step 6.* He concludes that the population mean body temperature of African elephants is not equal to 96.0°F.

Hypothesis testing for a population mean: Example

In this example, we implemented Steps 4a and 5a. Alternatively, we could have used the "P value approach" of Steps 4b and 5b.

The *P value* of a computed test statistic is the probability, under H_0, that if we repeated the experiment we would get a computed test statistic as extreme or more extreme than the one we got for the experiment we actually did.

- *Step 4b.* Computed test statistic $=$, so

$$P \text{ value} = P(t_{49} > \quad) + P(t_{49} < \quad) =$$

- *Step 5b.* Since P value $< \alpha$, we reject H_0.

Hypothesis testing: More on P values

- The P value approach to HT is equivalent to the critical value approach; they always yield the same decision about H_0.
- The P value is a measure of the strength of evidence against H_0 that the data provides: a smaller P value corresponds to stronger evidence against H_0.
- The P value may also be interpreted as the value of α at which we would go from rejecting H_0 to not rejecting H_0, if we repeatedly retested our hypotheses at significance levels starting at $\alpha = 1$ and decreasing α towards 0.

Hypothesis testing: Variations on the elephant example

Suppose we change the significance level from .05 to .01. How does the test change?

Hypothesis testing: Variations on the elephant example

Suppose we return to using $\alpha = .05$, but the investigator has reason to believe, or wants to show, that the population mean body temperature of African elephants is less than 96.0. Thus, he wants to test

$$H_0 : \mu \geq 96.0 \qquad H_a : \mu < 96.0.$$

How does the test change?

Equivalence between hypothesis tests and confidence intervals

Consider testing

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0$$

at the α level of significance (and suppose σ is unknown). Then we will reject H_0 if our test statistic t satisfies

$$t < -t_{1-\alpha/2,n-1} \qquad \text{or} \qquad t > t_{1-\alpha/2,n-1}.$$

Equivalently, we will not reject H_0 if

$$-t_{1-\alpha/2,n-1} \leq \frac{\overline{X} - \mu_0}{s/\sqrt{n}} \leq t_{1-\alpha/2,n-1}.$$

Equivalence between hypothesis tests and confidence intervals

Manipulating this last inequality to get μ_0 by itself in the middle, we obtain the equivalent inequality

$$\overline{X} - t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}} \leq \mu_0 \leq \overline{X} + t_{1-\alpha/2,n-1}\frac{s}{\sqrt{n}}.$$

But observe that the endpoints of this interval coincide with the endpoints of a $100(1-\alpha)\%$ confidence interval for μ!!

Thus, if μ_0 is a value inside the $100(1-\alpha)\%$ confidence interval for μ, we will not reject H_0; otherwise we will reject H_0.

Bottom line: If we don't care about reporting a P value, we can simply use the appropriate confidence interval to perform a two-sided test for a population mean.

Hypothesis testing for a population variance: Introduction

Sometimes we wish to test two competing hypotheses about a population variance (rather than a mean).

Example: As a result of the manufacturing process for a particular drug, there is some variation in the actual dosage of the active ingredient included in each pill. Suppose that the variance in actual dosage is known to be 0.36 μg^2. A new manufacturing process is developed which is believed to reduce this variance. Letting σ^2 represent the variance in the population of pills manufactured by the new process. The developer of the new process wants to test

$$H_0 : \sigma^2 \geq 0.36 \qquad \text{versus} \qquad H_a : \sigma^2 < 0.36.$$

Hypothesis testing for a population variance: Introduction

More generally, hypotheses about a population variance are of 3 types:

- H_0: $\sigma^2 = \sigma_0^2$ versus $H_a : \sigma^2 \neq \sigma_0^2$
- H_0: $\sigma^2 \leq \sigma_0^2$ versus $H_a : \sigma^2 > \sigma_0^2$
- H_0: $\sigma^2 \geq \sigma_0^2$ versus $H_a : \sigma^2 < \sigma_0^2$

Note that a hypothesis about a population standard deviation is equivalent to a hypothesis about a population variance, e.g.

$$H_0 : \sigma^2 = \sigma_0^2 \qquad \Leftrightarrow \qquad H_0 : \sigma = \sigma_0$$

HT for a population variance: Test statistic

Based on a random sample of size n from the population, our best estimate of σ^2 is the sample variance, s^2.

To measure the discrepancy between what the data say about the variance (best guess is s^2) and what the null hypothesis claims about the variance (claimed to equal σ_0^2), we could use

$$s^2 - \sigma_0^2.$$

However, it turns out that a better and more convenient measure of the discrepancy is the ratio, s^2/σ_0^2, and even more convenient is the scaled ratio,

$$\frac{(n-1)s^2}{\sigma_0^2}. \qquad \text{(book's denominator is } \sigma^2\text{)}$$

HT for a population variance: Test statistic

The further that s^2/σ_0^2 is from 1.0, the greater the discrepancy (and the stronger the evidence against H_0). Equivalently, the further that

$$\frac{(n-1)s^2}{\sigma_0^2}$$

is from $n-1$, the stronger the evidence is against H_0. So this is our test statistic.

For critical values and P values, we use the fact that our test statistic has a chi-square distribution with $n-1$ degrees of freedom when H_0 is true.

HT for a population variance: Critical values

Suppose the significance level is α. Let $\chi^2_{n-1,\alpha}$ be the 100αth percentile of the χ^2_{n-1} distribution, i.e.

$$P(\chi^2_{n-1} < \chi^2_{n-1,\alpha}) = \alpha.$$

Then:

- if H_a is two-sided, we reject H_0 if

- if H_a is $H_a : \sigma^2 > \sigma_0^2$, we reject H_0 if
- if H_a is $H_a : \sigma^2 < \sigma_0^2$, we reject H_0 if

HT for a population variance: P values

If we wish to use the P value approach to HT instead, then we compute the P value as follows:

- if H_a is two-sided, P value =

- if H_a is $H_a : \sigma^2 > \sigma_0^2$, P value =

- if H_a is $H_a : \sigma^2 < \sigma_0^2$, P value =

HT for a population variance: Example

A healthy lifestyle undoubtedly plays a role in longevity, but so does genetic makeup. Recent studies have linked large cholesterol particles to longevity. A variant of a gene called CETP encoding the cholestryl ester transferase protein apparently causes the formation of large cholesterol particles. In a particular population the life spans for males are normally distributed with a mean of 74.2 yrs and a standard deviation of 10.0 yrs. A sample of 16 males in this population that had the variant CETP gene lived an average of 81.2 yrs with a standard deviation of 8.0 yrs. Does this establish that CETP variant carriers are significantly less variable in their life spans than the general population?

Nonparametric methods for hypothesis testing

The HT methods we've learned so far require either (a) the sampled population to be normally distributed, or (b) the sample size to be large enough for the CLT to "steer" the distribution of the test statistic sufficiently close to its reference distribution (Z, t, χ^2). What if neither (a) nor (b) is satisfied?

In that case, we use alternative HT methods called *nonparametric* or *distribution-free* methods. Though more widely applicable than the *parametric* methods already learned, they are not as powerful.

The sign test: Introduction

The first nonparametric HT method we learn is the sign test.

- tests hypotheses on the median, M, of a continuous population
- based on the idea that if $M = M_0$, then roughly half of the observations in a random sample drawn from the population should be greater than M_0, and half less than M_0
- If too few observations in the sample lie to one side of M_0, that suggests that M is not equal to M_0.

The sign test: Hypotheses and test statistics

The hypotheses to be tested are one of three types:

- H_0: $M = M_0$ versus $H_a : M \neq M_0$
- H_0: $M \leq M_0$ versus $H_a : M > M_0$
- H_0: $M \geq M_0$ versus $H_a : M < M_0$

The test statistic is one or the other of:

$$\begin{aligned} S_- &= \text{\# of observations in sample less than } M_0, \\ S_+ &= \text{\# of observations in sample greater than } M_0. \end{aligned}$$

If any observations in the sample equal M_0, they are ignored; n excludes them also (so here n represents a possibly reduced sample size).

The sign test: Test statistics and reference distribution

Based on a random sample of size n, we compute S_- and S_+. Now, each of these is a binomial random variable; in fact, if $M = M_0$ then

So our reference distribution is bin$(n, \frac{1}{2})$. If tabled critical values were available for this distribution (for the commonly used levels of significance) we would use them, but they aren't. So we use a P value testing approach instead.

The sign test: P value testing approach

We use a P value approach to make our decision about the hypotheses. Three cases:

- If H_a is $M > M_0$, we reject H_0 if S_- is too small, i.e. if
$$P(\text{bin}(n, \tfrac{1}{2}) \leq S_-) < \alpha.$$
- If H_a is $M < M_0$, we reject H_0 if S_+ is too small, i.e. if
$$P(\text{bin}(n, \tfrac{1}{2}) \leq S_+) < \alpha.$$
- If H_a is two-sided, we reject H_0 if $\min(S_-, S_+)$ is too small, i.e. if
$$2P(\text{bin}(n, \tfrac{1}{2}) \leq \min(S_-, S_+)) < \alpha.$$

The sign test: P value testing approach

How do we get the binomial probabilities needed to do the test?

- If $5 \leq n \leq 20$, we can use the column corresponding to $p = 0.5$ in Table C.1 directly.
- If $n > 20$, we invoke the CLT to approximate the bin$(n, \frac{1}{2})$ distribution by the N$(\frac{n}{2}, \frac{n}{4})$ distribution:

$$P(\text{bin}(n, \tfrac{1}{2}) \leq S) \doteq P\left(Z \leq \frac{S + 0.5 - \frac{n}{2}}{\sqrt{n/4}}\right).$$

(Here S is either S_- or S_+, depending on which one is applicable.)

The sign test: Example 1 (Problem 6.32 in text)

Abuse of substances containing toluene (for example, various glues) can produce neurological symptoms. In an investigation of the mechanism of these toxic effects, researchers measured the concentrations of certain chemicals in the brains of rats who had been exposed to a toluene-laden atmosphere. The concentrations (ng/gm) of the brain chemical norepinephrine in the medulla region of the brain of 9 toluene-exposed rats was determined and recorded below:

543 523 431 635 564 580 600 610 550

Does the exposure to toluene significantly increase norepinephrine levels in rat medullas above the normal median level of 530 ng/gm?

The sign test: Example 2

According to the National Center for Health Statistics, the median height of adult women in the United States is 63.6 inches. Assuming that the women in our class are a random sample from the population of women at UI, is the median height of UI women different than this?

The Wilcoxon signed-rank test: Introduction

- Tests hypotheses about a population median, M (tests same hypotheses as sign test)
- Its use requires the population distribution to be symmetric, so it's not as widely applicable as the sign test
- Since symmetry implies that the mean and median are equal, this test tests the same hypotheses as the t test too
- Based on the idea that when the population distribution is symmetric and $M = M_0$, the distances of sampled observations from M_0 should be about the same on both sides of it
- More powerful than sign test; not as powerful as t test

The Wilcoxon signed-rank test: Test statistic

Follow these steps to compute the test statistic:

1. Form the differences, $X_i - M_0$, and take their absolute values to get the distances of observations from the hypothesized median, $|X_i - M_0|$.

2. Rank those distances from smallest to largest and replace the numerical distance with its rank.

3. Add a plus or minus sign to the rank according to whether the original difference was positive or negative.

4. Compute $W_+ =$ sum of the positive signed ranks, and $W_- = -1 \times$ sum of the negative signed ranks.

The Wilcoxon signed-rank test: Test statistic

Note:

- Once again we ignore any observations that equal M_0 (and reduce n accordingly).
- Also, if ties occur we average all the successive ranks that are tied. For example,

The Wilcoxon signed-rank test statistic: Example

Recall the example of toluene-exposed rats used to illustrate the sign test. The data (epinephrine concentrations in medulla) and the steps of the test statistic computation are as follows:

X_i	543	523	431	635	564	580	600	610	550
$X_i - 530$	13	-7	-99	105	34	50	70	80	20
$\lvert X_i - 530 \rvert$	13	7	99	105	34	50	70	80	20
Rank	2	1	8	9	4	5	6	7	3
Signed rank	2	-1	-8	9	4	5	6	7	3

$$W_+ = 36, \qquad W_- = 9$$

The Wilcoxon signed-rank test: Rationale

Recall the following fact:

$$1+2+3+\cdots+n=\frac{n(n+1)}{2}.$$

So, if the population distribution is symmetric and $M=M_0$, we would expect both W_+ and W_- to equal roughly $n(n+1)/4$. If either of them is too small (too far away from $n(n+1)/4$), we should reject H_0.

How far is too far? We need a reference distribution for the Wilcoxon signed rank statistic, which is provided in Table C.6. Call this distribution $W(n)$.

The Wilcoxon signed-rank test: P value testing approach

Three cases:

- If H_a is $M > M_0$, we reject H_0 if W_- is too small, i.e. if

$$P(W(n) \leq W_-) < \alpha.$$

- If H_a is $M < M_0$, we reject H_0 if W_+ is too small, i.e. if

$$P(W(n) \leq W_+) < \alpha.$$

- If H_a is two-sided, we reject H_0 if $\min(W_-, W_+)$ is too small, i.e. if

$$2P(W(n) \leq \min(W_-, W_+)) < \alpha.$$

The Wilcoxon signed-rank test: P value testing approach

To get the probabilities for the aforementioned approach:

- If $5 \leq n \leq 25$, we can use Table C.6 directly.
- If $n > 25$, we invoke the CLT to approximate the $W(n)$ distribution by the N($\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}$) distribution:

$$P(W(n) \leq w) \doteq P\left(Z \leq \frac{w + 0.5 - \frac{n(n+1)}{4}}{\sqrt{n(n+1)(2n+1)/24}}\right).$$

(Here w is either W_- or W_+, depending on which one is applicable.)

The Wilcoxon signed-rank test: Example 1

Let's finish up the example of the epinephrine concentrations in the medullas of toluene-exposed rats. Recall that $W_+ = 36$ and $W_- = 9$. Since H_a is $M > 530$, the P value is

The Wilcoxon signed-rank test: Example 2

According to the National Center for Health Statistics, the median height of adult women in the United States is 63.6 inches. Assuming that the women in our class are a random sample from the population of women at UI, is the median height of UI women different than this?

For these data, $n =$, $W_+ =$, W_- .

$P(W(n) \leq$ $) =$

Comparing two population means: Introduction

In some situations, the scientific question of interest is not so much what the mean of a single population is, but how the means of two distinct populations compare to each other.

Example: The ages (in days) at time of death for random samples of 11 girls and 16 boys who died from SIDS were as follows:

Girls				53	56	60	60	78	87	102	117	134	160	277			
Boys	46	52	58	59	77	78	80	81	84	103	114	115	133	134	167	175	

The question of interest: Are the mean ages at death due to SIDS identical for boys and girls?

Comparing two population means: Introduction

The previous example is of the following general type.

Suppose there are two populations of interest, the first with mean μ_1 and variance σ_1^2, say, and the second with mean μ_2 and variance σ_2^2. From the first population we take a random sample of size n_1:

$$X_{11}, X_{12}, \ldots, X_{1n_1}.$$

From the second population we take a random sample of size n_2:

$$X_{21}, X_{22}, \ldots, X_{2n_2}.$$

Based on the information in these two samples, we wish to estimate the difference in the two means, i.e. $\mu_1 - \mu_2$, and perhaps test hypotheses about this difference.

Comparing two population means: Types of hypotheses

Hypotheses comparing two population means are of 3 types:

- H_0: $\mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$
- H_0: $\mu_1 \leq \mu_2$ versus $H_a : \mu_1 > \mu_2$
- H_0: $\mu_1 \geq \mu_2$ versus $H_a : \mu_1 < \mu_2$

Note that each hypothesis can also be expressed in terms of the difference of the means. For example, the last pair of hypotheses can also be expressed as follows:

- H_0: $\mu_1 - \mu_2 \geq 0$ versus $H_a : \mu_1 - \mu_2 < 0$

Comparing two population means: Types of sampling

The random samples can be of two distinct types:

1. *Paired samples* — for each i, the ith individual in the first sample is more closely related, in some meaningful way, to the ith individual in the second sample than it is to the other individuals in the second sample.

2. *Independent samples* — for each i, the ith individual in the first sample is no more closely related, in any meaningful way, to any individual in the second sample than it is to other individuals in the second sample.

For paired samples, $n_1 = n_2$; not necessarily so for independent samples.

Paired versus independent sampling: Examples

- Bacterial contamination in meat samples, before and after irradiation

- Weight gain for cattle on two feeds

- Studies of survival after two disease therapies

Comparing two means via paired sampling

When sampling is paired, the data can be listed in pairs:

$$(X_{11}, X_{21}), (X_{12}, X_{22}), \ldots, (X_{1n}, X_{2n})$$

and we can reduce the data to a single sample of within-pair differences

$$d_i = X_{1i} - X_{2i}, \quad i = 1, \ldots, n.$$

(Here we write n for n_1 and n_2.)

Reducing the data to within-pair differences controls for sources of variation in the data other than the source of main interest (more on this later).

Comparing two means via paired sampling

In this context we let μ_d represent $\mu_1 - \mu_2$; equivalently, μ_d is the mean of the population of within-pair differences. The d_i's are a random sample from this population.

The natural point estimate of μ_d is

$$\overline{X}_d = \frac{1}{n}\sum_{i=1}^{n} d_i,$$

and the natural estimate of the variance of the population of within-pair differences is

$$s_d^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} d_i^2 - \frac{(\sum_{i=1}^{n} d_i)^2}{n}\right).$$

Comparing two means via paired sampling

A $100(1-\alpha)\%$ confidence interval for μ_d is

$$\overline{X}_d \pm t_{1-\alpha/2,n-1}\frac{s_d}{\sqrt{n}}.$$

Also, to test hypotheses about μ_d, we use the same test statistic, same critical value(s), same everything used for HT about the mean μ of a single population — but applied here to the within-pair differences. For example, to test $H_0 : \mu_d = 0$ versus $H_a : \mu_d \neq 0$ at the α sigificance level, we reject H_0 if

$$\frac{\overline{X}_d - 0}{s_d/\sqrt{n}} < -t_{1-\alpha/2,n-1} \qquad \text{or} \qquad \frac{\overline{X}_d - 0}{s_d/\sqrt{n}} > t_{1-\alpha/2,n-1}.$$

Comparing two means via paired sampling: Example

An experiment was performed to study the effects of irradiation on bacterial contamination in meat. The logarithm of the direct microscopic count (log DMC) of bacteria in 12 meat samples was measured before irradiating the 12 meat samples, and then again afterwards. The data were as follows:

log DMC, before	log DMC, after	d_i (before minus after)
6.98	6.95	.03
7.08	6.94	.14
8.34	7.17	1.17
5.30	5.15	.15
6.26	6.28	–.02
6.77	6.81	–.04
7.03	6.59	.44
5.56	5.34	.22
5.97	5.98	–.01
6.64	6.51	.13
7.03	6.84	.19
7.69	6.99	.70

$$\overline{X}_d = .258, \qquad s_d^2 = .127$$

Comparing two means via paired sampling: Example

The investigator wanted to show that irradiation reduces bacterial contamination so, letting μ_d represent the mean change (before minus after) in log DMC due to irradiation for the conceptual population of all possible meat samples, the hypotheses of interest are

$$H_0 : \mu_d \leq 0 \qquad \text{versus} \qquad H_a : \mu_d > 0.$$

The computed test statistic is

$$\frac{\overline{X}_d - 0}{s_d/\sqrt{n}} = \frac{.258}{\sqrt{.127/12}} = 2.51.$$

If we test at the .05 significance level, the critical value is $t_{.95,11} = 1.796$, so we reject H_0. Conclusion: there is statistically significant evidence that irradiation reduces bacterial contamination in meat.

Comparing two means via independent sampling

For data arising from independent sampling of two populations, we use the data as they are (we don't form differences).

Our best point estimate of $\mu_1 - \mu_2$ is $\overline{X}_1 - \overline{X}_2$, but how we use this to obtain CI's and do HT's depends on what is assumed about the population variances (σ_1^2 and σ_2^2). Two cases:

1. Assume $\sigma_1^2 = \sigma_2^2$
2. Do not assume $\sigma_1^2 = \sigma_2^2$

Comparing two means via independent sampling, assuming equal variances

Assuming that $\sigma_1^2 = \sigma_2^2$, the two sample variances (s_1^2 and s_2^2) are estimates of the same quantity. So it makes sense to combine, or "pool" them:

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

- The divisor, $n_1 + n_2 - 2$, is the degrees of freedom here
- The sample variance from the larger of the two samples gets more weight in the pooled estimate (makes sense!)
- If $n_1 = n_2$, then s_p^2 is merely the average of the two sample variances

Comparing two means via independent sampling, assuming equal variances

A $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\alpha/2, n_1+n_2-2} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}.$$

Furthermore, to test hypotheses comparing the two population means, we use

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}}$$

as our test statistic.

Comparing two means via independent sampling, assuming equal variances

More specifically:

- To test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$, we reject H_0 if

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} < -t_{1-\alpha/2, n_1+n_2-2}$$

or

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} > t_{1-\alpha/2, n_1+n_2-2}.$$

Comparing two means via independent sampling, assuming equal variances

- To test $H_0 : \mu_1 \geq \mu_2$ versus $H_a : \mu_1 < \mu_2$, we reject H_0 if

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} < -t_{1-\alpha, n_1+n_2-2}$$

- To test $H_0 : \mu_1 \leq \mu_2$ versus $H_a : \mu_1 > \mu_2$, we reject H_0 if

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2(\frac{1}{n_1} + \frac{1}{n_2})}} > t_{1-\alpha, n_1+n_2-2}$$

Comparing two means via independent sampling, assuming equal variances: Example

In a study of the periodical cicada (*Magicicada septendecim*), researchers measured the hind tibia lengths of the shed skins of 110 individuals: 60 males and 50 females. Some summary statistics for the tibia lengths were as follows:

Gender	n_i	$\overline{X}_i$	s_i
Males	60	78.42	2.87
Females	50	80.44	3.52

Let μ_1 and σ_1^2 represent the mean and variance of hind tibia lengths for the entire population of male periodical cicadas at shedding; define μ_2 and σ_2^2 similarly for females. We want to test $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ at, say, the .05 level of significance, and suppose we're willing to assume that $\sigma_1^2 = \sigma_2^2$.

Comparing two means via independent sampling, assuming equal variances: Example

Pooled sample variance:

Test statistic:

Critical values:

P-value:

Conclusion:

Comparing two means via independent sampling, assuming equal variances: Example

95% confidence interval for $\mu_1 - \mu_2$:

95% confidence interval for μ_1:

95% confidence interval for μ_2:

Comparing two means via independent sampling, when variances are possibly unequal

In this case s_1^2 and s_2^2 aren't necessarily estimating the same quantity, so we do not pool them. We sum them instead (actually we sum scaled versions of them).

$100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$:

$$(\overline{X}_1 - \overline{X}_2) \pm t_{1-\alpha/2,\nu} \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

The degrees of freedom here, represented by the symbol ν, is determined using a messy expression given on page 196 of the text.

Comparing two means via independent sampling, when variances are possibly unequal

To test hypotheses comparing the means, we follow the procedure for the equal variances case described on pp. 234-235 of these notes, with two important differences:

1. Replace the test statistic there with

$$\frac{\overline{X}_1 - \overline{X}_2}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

2. Replace the df there (which is $n_1 + n_2 - 2$) with ν

Comparing two means via independent sampling, when variances are possibly unequal: Example

Let's revisit the periodical cicada example, only this time let's not assume that the two population variances are equal. Then our test statistic is

The df, ν, is:

So critical values are $\pm t_{.975,94} = \pm 1.986$. We still reject H_0, but slightly less emphatically (test statistic not as extreme, critical values more extreme).

Pros and cons of paired sampling

Paired sampling is to be preferred over independent sampling when the within-pair differences are likely to be less variable than the data from individuals in each sample.

Why? Because when the variability of within-pair differences is less than the variability among individuals in the same sample, paired-sampling based confidence intervals for the difference in the two population means will be narrower, and paired-sampling based hypothesis tests about the two population means will be more powerful.

Why? Consider the following numerical demonstrations. For both, suppose we take paired samples of size 5 from two populations.

Pros and cons of paired sampling: Numerical demonstration

Case #1: Pairing is effective

First sample	Second sample	Within-pair difference
1	0	1
6	5	1
10	10	0
17	15	2
21	20	1

$$\overline{X}_d = 1, \qquad s_d^2 = (0+0+1+1+0)/4 = 0.5.$$

95% confidence interval for μ_d:

$$\overline{X}_d \pm t_{.975,4} s_d/\sqrt{n} = 1 \pm 2.776\sqrt{0.5}/\sqrt{5} = 1 \pm 0.88.$$

The t-test rejects $H_0 : \mu_d = 0$ (against a two-sided H_a) at .05 significance level (test statistic = 3.16, critical values are ± 2.776).

Pros and cons of paired sampling: Numerical demonstration

Suppose we do an independent-sampling based analysis of the same data:

$$\overline{X}_1 - \overline{X}_2 = 1, \qquad s_1^2 = (100+25+1+36+100)/4 = 65.5,$$

$$s_2^2 = (100+25+0+25+100)/4 = 62.5,$$

$$s_p^2 = \frac{4(62.5)+4(65.5)}{8} = 64.$$

A 95% confidence interval for $\mu_1 - \mu_2$ is

$$(\overline{X}_1 - \overline{X}_2) \pm t_{.975,8} s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}} = 1 \pm 2.306(8)\sqrt{2/5} = 1 \pm 11.67.$$

The t-test does not reject $H_0 : \mu_1 = \mu_2$ (against a two-sided H_a) at the .05 significance level (test statistic = 0.197, critical values are ± 2.306).

Thus, in this case the pairing was very effective, and the paired-sampling analysis leads to a much better (narrower) confidence interval for the difference in means, and a much more powerful test.

Pros and cons of paired sampling: Numerical demonstration

Case #2: Pairing is ineffective

First sample	Second sample	Within-pair difference
21	0	21
17	5	12
10	10	0
6	15	-9
1	20	-19

$$\overline{X}_d = 1, \qquad s_d^2 = (400+121+1+100+400)/4 = 255.5.$$

95% confidence interval for μ_d:

$$\overline{X}_d \pm t_{.975,4} s_d/\sqrt{n} = 1 \pm 2.776\sqrt{255.5}/\sqrt{5} = 1 \pm 19.84.$$

The t-test does not reject $H_0 : \mu_d = 0$ (against a two-sided H_a) at .05 significance level (test statistic = 0.14, critical values are ± 2.776).

Pros and cons of paired sampling: Numerical demonstration

Note that an independent-sampling based analysis of the second set of data is identical to an independent-sampling based analysis of the first set of data.

Thus, in Case #2 the paired-sampling based analysis is even worse than the independent-sampling based analysis.

Moral of the story: To be effective, a paired-sampling based analysis must be based on effective paired sampling. For paired sampling to be effective, it must remove (control for) at least some variation between individuals, so that the variability of differences within pairs is less than the variability among individuals in each sample.

Sources of variation that paired sampling controls for: Examples

- Bacterial contamination in paired meat samples, before and after irradiation

- Weight gain for twin calves on two feeds

- Studies of survival of "matched pairs" after two disease therapies

The irradiated meat example, revisited: Incorrectly analyzed by an independent-sampling based approach

Although the sampling was paired in the irradiated meat example, let's see what happens when we incorrectly act as though the sampling was independent. For simplicity assume that $\sigma_1^2 = \sigma_2^2$.

Relevant summary statistics:

$$n_1 = n_2 = 12, \qquad \overline{X}_1 = 6.721, \qquad \overline{X}_2 = 6.463,$$

$$s_1^2 = 0.6565, \qquad s_2^2 = 0.4341, \qquad s_p^2 = 0.5453,$$

$$\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)} = 0.301.$$

For testing $H_0 : \mu_1 \leq \mu_2$ vs. $H_a : \mu_1 > \mu_2$ at the .05 significance level, the critical value of t is $t_{.95,22} = 1.717$. The test statistic is

$$t = \frac{\overline{X}_1 - \overline{X}_2}{\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} = 0.857.$$

So, based on this analysis, we do not reject H_0 — a different conclusion from that reached by the paired-sampling based analysis.

Why the difference? Less variability among within-pair differences than among the data in each sample (compare denominators of the test statistics:

$$\sqrt{s_p^2 \left(\frac{1}{n_1} + \frac{1}{n_2}\right)} = 0.301 \qquad \text{vs.} \qquad s_d/\sqrt{n} = 0.103.)$$

Nonparametric tests for two populations: Introduction

The previously described tests for the means of two populations are strictly valid only when either:

- the population of within-pair differences (for paired sampling) or the two sampled populations themselves (for independent sampling) are normally distributed; or
- the sample size n (for paired sampling), or both of n_1 and n_2 (for independent sampling), are large enough for the CLT to apply.

If neither of these holds, then we can instead use nonparametric tests.

Nonparametric tests for two populations: Introduction

When the sampling is paired, we may test hypotheses about the median of the population of within-paired differences using either the

- sign test, or
- Wilcoxon signed-rank test (if the population is symmetric).

When the sampling is independent, we may test hypotheses about the medians of the two populations using a test called the Wilcoxon rank-sum test.

The irradiated meat example, revisited: Sign test and Wilcoxon signed-rank test for the population median of paired differences

Recall, once again, the following log DMC data in 12 meat samples before and after irradiation. The investigator wishes to test

$$H_0 : M_d \leq 0 \quad \text{versus} \quad H_a : M_d > 0.$$

log DMC, before	log DMC, after	d_i
6.98	6.95	.03
7.08	6.94	.14
8.34	7.17	1.17
5.30	5.15	.15
6.26	6.28	–.02
6.77	6.81	–.04
7.03	6.59	.44
5.56	5.34	.22
5.97	5.98	–.01
6.64	6.51	.13
7.03	6.84	.19
7.69	6.99	.70

For these data,

$$S_- = \quad , \qquad S_+ = \quad , \qquad W_- = 7, \qquad W_+ = 71.$$

P-value for sign test:

P-value for Wilcoxon signed rank test:

Thus, the sign test does not reject H_0 (at $\alpha = .05$), but the Wilcoxon signed rank test does reject H_0 (indicating that there is and is not, respectively, statistically significant evidence that irradiation reduces bacterial contamination in meat).

Why the difference in conclusions?

Wilcoxon rank-sum test

When we wish to compare two population medians, and the sampling of those populations is independent, we may use the Wilcoxon rank-sum test, if we are willing to assume that the two populations have the same shape.

The hypotheses to be tested are one of the following:

- H_0: $M_1 = M_2$ versus $H_a : M_1 \neq M_2$
- H_0: $M_1 \leq M_2$ versus $H_a : M_1 > M_2$
- H_0: $M_1 \geq M_2$ versus $H_a : M_1 < M_2$

Wilcoxon rank-sum test: Test statistic

Important: Label your samples (and populations) such that the smaller of the two sample sizes is n_1 (if the two samples are equal in size, then it doesn't matter how you label them).

Procedure for computing the test statistic:

1. Conceptually pool the data from both samples into one sample and rank the data from smallest to largest. Replace the data with their ranks.

2. Sum the ranks that correspond to the smaller of the two samples (Sample 1). Call this rank sum W_1, which is our test statistic.

Wilcoxon rank-sum test: Critical values

Critical values are listed in Table C.8:

- Book's m and n are my n_1 and n_2, respectively; the table gives critical values for sample sizes in the range $3 \leq n_1 \leq n_2 \leq 25$.
- The significance level options in the table are very limited.
- Column labeled W gives the lower and upper critical values. Use both critical values for a two-sided test; use only the lower one if H_a is $M_1 < M_2$; use only the upper one if H_a is $M_1 > M_2$.
- Column labeled P gives the P value for a one-sided test whose test statistic exactly equals the critical value; this rarely happens so you can ignore it.

Wilcoxon rank-sum test: Example

Recall (from page 221 of these notes) the following data, which are the ages (in days) at time of death for random samples of 11 girls and 16 boys who died from SIDS:

Girls				53	56	60	60	78	87	102	117	134	160	277			
Boys	46	52	58	59	77	78	80	81	84	103	114	115	133	134	167	175	

Histograms of these data show that for both girls and boys, the data are right-skewed. Thus, the corresponding populations are probably not normally distributed, and the sample sizes are relatively small.

Question of interest: Are the median ages at death due to SIDS identical for boys and girls?

So we wish to test

$$H_0 : M_1 = M_2 \qquad \text{versus} \qquad H_a : M_1 \neq M_2.$$

Original data:

Girls				53	56	60	60	78	87	102	117	134	160	277			
Boys	46	52	58	59	77	78	80	81	84	103	114	115	133	134	167	175	

Ranks in pooled sample:

Girls				3	4	7.5	7.5	10.5	15	16	20	22.5	24	27			
Boys	1	2	5	6	9	10.5	12	13	14	17	18	19	21	22.5	25	26	

Test statistic: $W_1 = 157$.
Critical values for two-sided test at $\alpha = .05$: 113 and 195.
Since our test statistic is not as extreme as the critical values, we do not reject H_0. Scientific conclusion: Among those babies that die of SIDS, there is no significant difference between male's and female's median ages at death.

Comparing two population variances: Introduction

Several pages back, we saw that the specifics of inference for the difference in two population means based on independent samples depended on whether or not we assumed that the two population variances were equal. If we do make this assumption, it's desirable to have some justification for doing so. This leads us to consider the problem of testing the hypothesis that the two population variances are equal.

We may also be interested in testing whether two population variances are equal for its own sake.

Comparing two population variances: Hypotheses

The hypotheses to be tested are one of the following:

- H_0: $\sigma_1^2 = \sigma_2^2$ versus $H_a : \sigma_1^2 \neq \sigma_2^2$
- H_0: $\sigma_1^2 \leq \sigma_2^2$ versus $H_a : \sigma_1^2 > \sigma_2^2$
- H_0: $\sigma_1^2 \geq \sigma_2^2$ versus $H_a : \sigma_1^2 < \sigma_2^2$

Observe that we can re-express any of these in terms of the ratio of the two variances. For example, the third pair of hypotheses can be rewritten as

- H_0: $\sigma_1^2/\sigma_2^2 \geq 1$ versus $H_a : \sigma_1^2/\sigma_2^2 < 1$

Comparing two population variances: Test statistic

As a test statistic, consider the ratio of the two sample variances, which we label as F:

$$F = \frac{s_1^2}{s_2^2}$$

If the two population variances are equal (equivalently, if $\sigma_1^2/\sigma_2^2 = 1$), then F should usually be pretty close to 1. Extreme (too large or too small) values of F would cast doubt on the hypothesis that the two population variances are equal.

What do we compare F to, to determine if it's extreme enough to reject H_0?

Comparing two population variances: Critical values

When H_0 is true, F has a well-known (and tabled) distribution called the *F distribution*, so it is this distribution that we refer to to obtain critical values (and P values).

The F distribution:

- has a shape similar to the chi-square distribution (unimodal, bell-shaped, right-skewed, positive pdf only over positive half-line)
- is really a family of distributions — one for each of two degree-of-freedom parameters, ν_1 and ν_2, called the numerator df and denominator df
- is tabled in Table C.7

Comparing two population variances: Critical values

Values in Table C.7 are critical values in the right tail, i.e. $F_{1-\alpha,(\nu_1,\nu_2)}$:

- $F_{.95,(6,11)} =$
- $F_{.99,(60,25)} =$
- $F_{.975,(47,83)} \doteq$

Critical values in the left tail are not tabled (to save space), but can be obtained from critical values in the right tail as follows:

$$F_{\alpha,(\nu_1,\nu_2)} = 1/F_{1-\alpha,(\nu_2,\nu_1)}.$$

(Note that the order of df are reversed on the RHS.)

Comparing two population variances: Critical values

For comparing two population variances,

$$\nu_1 = n_1 - 1 \qquad \text{and} \qquad \nu_2 = n_2 - 1.$$

So for the F-test, critical values are:

- $1/F_{1-\alpha/2,(n_2-1,n_1-1)}$ and $F_{1-\alpha/2,(n_1-1,n_2-1)}$, if H_a is two-sided
- $F_{1-\alpha,(n_1-1,n_2-1)}$, if H_a is $\sigma_1^2 > \sigma_2^2$
- $1/F_{1-\alpha,(n_2-1,n_1-1)}$, if H_a is $\sigma_1^2 < \sigma_2^2$

If our F test statistic is more extreme than the critical value(s), we reject H_0; otherwise, we do not reject H_0.

Comparing two population variances: Example

Let's revisit the periodical cicada example (from p. 236 of these notes) to see if the assumption of equal population variances that we made there can be justified. We wish to test

$$H_0 : \sigma_1^2 = \sigma_2^2 \qquad \text{versus} \qquad H_a : \sigma_1^2 \neq \sigma_2^2.$$

Relevant summary statistics:

$$n_1 = 60, \quad n_2 = 50, \quad s_1 = 2.87, \quad s_2 = 3.52$$

Test statistic: $F =$

Critical values (using $\alpha = .05$):

Conclusion:

Hypothesis testing: Some loose ends ...

1. Point estimate(s) satisfies null hypothesis

For a particular population, suppose we wish to test

$$H_0 : \mu \leq 65 \quad \text{versus} \quad H_a : \mu > 65.$$

Furthermore, suppose a random sample is taken from this population and the sample mean ($\overline{X}$) is 61. Then the test statistic is

In which tail is the critical value?

So we will not reject H_0 (at *any* typical level of significance).

Hypothesis testing: Some loose ends . . . (continued)

Similarly, for a situation with two populations, suppose we wish to test

$$H_0 : \mu_1 \geq \mu_2 \quad \text{versus} \quad H_a : \mu_1 < \mu_2.$$

If independent random samples are taken from the two populations, and the sample means satisfy $\overline{X}_1 \geq \overline{X}_2$, then the test statistic is

In which tail is the critical value?

So we will not reject H_0.

General Fact: When doing a one-sided test, if the point estimate satisfies H_0, we can skip calculation of the test statistic and conclude immediately that H_0 is not rejected (at typical levels of significance).

Hypothesis testing: Some loose ends . . . (continued)

2. Comparison of conclusion for one-sided and two-sided alternatives

(a) Suppose we are asked to test

$$H_0 : \mu \leq 65 \quad \text{versus} \quad H_a : \mu > 65$$

but we mistakenly test

$$H_0 : \mu = 65 \quad \text{versus} \quad H_a : \mu \neq 65.$$

Suppose further that $\overline{X}$ is larger than 65, and that we reject the second H_0 (at a particular level of significance, say .05).

Would we also reject the first H_0 (at the same level of significance)?

Hypothesis testing: Some loose ends . . . (continued)

(b) Same scenario as (a), except suppose that we do not reject the second H_0. Would we also not reject the first H_0?

Hypothesis testing: Some loose ends . . . (continued)

(c) Suppose we are asked to test

$$H_0 : \mu = 65 \quad \text{versus} \quad H_a : \mu \neq 65$$

but we mistakenly test

$$H_0 : \mu \leq 65 \quad \text{versus} \quad H_a : \mu > 65.$$

Suppose further that $\overline{X}$ is larger than 65, and that we reject the second H_0 (at a particular level of significance, say .05).

Would we also reject the first H_0 (at the same level of significance)?

Hypothesis testing: Some loose ends . . . (continued)

(d) Same scenario as (c), except suppose that we do not reject the second H_0. Would we also not reject the first H_0?

General fact: It takes a more extreme value of the test statistic to reject H_0 when H_a is two-sided than it does to reject H_0 when H_a is one-sided (at the same level of significance).

Hypothesis testing: Some loose ends . . . (continued)

3. Equivalence between hypothesis tests and confidence intervals

Recall, from pp. 195-196, that we can carry out the hypothesis test of

$$H_0 : \mu = \mu_0 \qquad \text{versus} \qquad H_a : \mu \neq \mu_0$$

at the α level of significance by rejecting H_0 if, and only if, μ_0 is <u>not</u> contained in the $100(1-\alpha)\%$ confidence interval for μ.

Likewise, we can carry out the hypothesis test of

$$H_0 : \mu_1 = \mu_2 \qquad \text{versus} \qquad H_a : \mu_1 \neq \mu_2$$

at the α level of significance by rejecting H_0 if, and only if, 0 is <u>not</u> contained in the $100(1-\alpha)\%$ confidence interval for $\mu_1 - \mu_2$.

Hypothesis testing: Some loose ends . . . (continued)

Similar equivalences can be stated for tests of other parameters (but don't worry about the details).

Testing hypotheses on more than two population means: Motivating example

Problem 8.8 from textbook, adapted from the article "Phytoestrogen supplements for the treatment of hot flashes: The isoflavone clover extract (ICE) study," *Journal of the American Medical Association*, vol. 290, pp. 207-214,by J. Tice et al., 2002:

The primary reason women seek medical attention for menopausal symptoms is hot flashes. Dietary supplements containing isoflavones derived from natural sources such as soy or red clover are marketed as an alternative treatment for such symptoms and are being used increasingly by women in the U.S. Isoflavones are polyphenol compounds that are similar in structure to estrogens.

A study was carried out to determine whether two dietary supplements derived from red clover were more effective than a placebo in reducing hot flashes in post-menopausal women. The randomized, double-blind trial was conducted using 252 menopausal women, aged 45 to 60 years, who were experiencing at least 35 hot flashes per week. After a 2-week period in which all were given a placebo, the women were randomly assigned to Promensil (82 mg of total isoflavones per day), Rimostil (57 mg of total isoflavones per day), or an identical placebo; and then followed up for 12 weeks. The table below provides summary statistics on the number of hot flashes (per day) experienced by the women at the end of the trial.

Testing hypotheses on more than two population means: Motivating example

	Promensil	Rimostil	Placebo
n_i	84	83	85
$\overline{X}_i$	5.1	5.4	5.0
s_i	4.1	4.6	3.2

Assuming normality, analyze these data to determine whether there are any differences in the mean number of hot flashes per day for these three treatments.

Hypotheses to test, using a two-population approach repeatedly:

Testing hypotheses on more than two population means: Motivating example

Suppose we test each hypothesis (using a two-sample t test, assuming equal population variances) at the .05 level of significance. What is the *overall* Type I error probability, i.e. the probability that we reject at least one of these null hypotheses when it is true?

Analogy to a monkey taking a three-question multiple-choice statistics exam, where each question has 20 choices:

Testing hypotheses on more than two population means: Motivating example

If the data being used to perform each hypothesis test was independent of the data being used to perform the others, the overall Type I error probability could be computed as follows:

But some of the data is the same in each of the tests, so independence doesn't hold and the previous calculation isn't valid.

Bottom line: We can't control (determine) the overall Type I error probability by doing multiple two-sample t tests. If we want to control α, we need to take a completely different approach, which is called the *Analysis of Variance* (ANOVA).

The ANOVA: Set-up and notation

So suppose we want to compare the means of k populations, where $k \geq 2$. (We already know how to do this when $k = 2$, so it's really $k > 2$ that interests us now.) We take independent random samples from each of the k populations.

Notation:

- μ_i: population mean of ith population
- σ_i^2: population variance of ith population
- n_i: size of sample from ith population
- N: total number of observations, i.e. $N =$

The ANOVA: Set-up and notation

More notation:

- X_{ij}: the jth observation in the ith sample
- $\overline{X}_{i.}$: sample mean of ith sample, i.e.,

$$\overline{X}_{i.} =$$

- $\overline{X}_{..}$: grand mean, i.e.

$$\overline{X}_{..} =$$

- s_i^2: sample variance of ith sample, i.e.,

$$s_i^2 =$$

The ANOVA: The hypotheses tested

The objective of an ANOVA is to test the null hypothesis

$$H_0 : \mu_1 = \mu_2 = \cdots = \mu_k$$

against the alternative hypothesis

$$H_a : \text{at least one } \mu_i \text{ is different from the others}$$

in such a way that the overall Type I error probability can be pre-specified (see pp. 277-278 of these notes).

The ANOVA: Underlying assumptions

1. The samples are independent random samples from their respective populations.

2. The populations are normally distributed, or if not, then sample sizes are large enough for the CLT to apply (all $n_i > 25$).

3. Population variances are all equal, i.e. $\sigma_1^2 = \sigma_2^2 = \cdots = \sigma_k^2$.

Sketch of population distributions satisfying the last 2 assumptions:

The ANOVA: Partitioning deviations from grand mean

Consider the deviation of an arbitrary observation from the grand mean, i.e.

$$X_{ij} - \overline{X}_{..}$$

By adding and subtracting the ith sample mean to this, we have the algebraic identity

$$X_{ij} - \overline{X}_{..} =$$

This partitions the deviation of an observation from the grand mean into two parts:

- the deviation of the group mean from the grand mean
- the deviation of the observation from its group mean

The ANOVA: Partitioning deviations from grand mean

Consider a toy example in which $k = 3$ and the data are as follows (lines separate the 3 samples, whose sample sizes are 3, 2, and 3, respectively):

Data	$X_{ij} - \overline{X}_{..}$	$\overline{X}_{i.} - \overline{X}_{..}$	$X_{ij} - \overline{X}_{i.}$
9	-11	-10	-1
10	-10	-10	
11	-9	-10	
19			
21			
28			
30			
32			

The ANOVA: Partitioning the sums of squares

Now let us square the deviations of observations from the grand mean, and then sum them up, i.e.

$$\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{..})^2.$$

It turns out that

$$\begin{aligned}\sum_{i=1}^{k}\sum_{j=1}^{n_i}(X_{ij}-\overline{X}_{..})^2 &= \sum_{i=1}^{k}\sum_{j=1}^{n_i}[(\overline{X}_{i.}-\overline{X}_{..})+(X_{ij}-\overline{X}_{i.})]^2 \\ &= \end{aligned}$$

(algebra details provided on p. 236 of textbook — never mind!).

We rewrite this as

$$\mathrm{SS}_{Total} = \mathrm{SS}_{Treat} + \mathrm{SS}_{Error}.$$

The ANOVA: Partitioning the sums of squares

Let's try this on our toy example:

$$\begin{aligned} \text{SS}_{Total} &= (-11)^2 + (-10)^2 + \cdots + (12)^2 = \\ \text{SS}_{Treat} &= (-10)^2 + (-10)^2 + \cdots + (10)^2 = \\ \text{SS}_{Error} &= (-1)^2 + (0)^2 + \cdots + (2)^2 = \end{aligned}$$

It works!

So what?

The ANOVA: Test statistic

All of the preceding algebraic development was for the purpose of computing a test statistic for testing the equality of means hypothesis described previously.

The test statistic is

$$F = \frac{\mathrm{SS}_{Treat}/(k-1)}{\mathrm{SS}_{Error}/(N-k)}.$$

We compare this to a right-tail critical value from the F distribution with $k-1$ and $N-k$ degrees of freedom. If $F > F_{1-\alpha,k-1,N-k}$, then we reject H_0; otherwise we do not reject H_0.

The ANOVA: Toy example

For our toy example,

$$\text{SS}_{Treat} = 600 \qquad \text{and} \qquad \text{SS}_{Error} = 12.$$

So,

$$F = \frac{600/(3-1)}{12/(8-3)} = 125.0$$

The critical value for a test at the .05 significance level is $F_{.95,2,5} = 5.79$, so we would reject H_0 for these "data."

The ANOVA: Some remarks

- Although the hypotheses being tested are concerned with population means, the test statistic is a ratio of measures of spread! Why is this reasonable?

- The alternative hypothesis is not one-sided (actually it's "multi-sided"), despite the fact that we reject H_0 only for values of F that are too large.

- There are better computational formulas for the sums of squares (see next two pages).

The ANOVA: Computational formulas for sums of squares

Additional notation:

- $T_{i.}$: sum of the observations in the ith sample, i.e.,

$$T_{i.} = \sum_{j=1}^{n_i} X_{ij}$$

- $T_{..}$: sum of all observations (over all samples), i.e.

$$T_{..} =$$

The ANOVA: Computational formulas for sums of squares

In practice, the following formulas for the sums of squares are algebraically equivalent, but not as painful or prone to numerical errors:

$$\begin{aligned}
\mathrm{SS}_{Total} &= \left(\sum_{i=1}^{k}\sum_{j=1}^{n_i} X_{ij}^2\right) - \frac{T_{..}^2}{N} \\
\mathrm{SS}_{Treat} &= \left(\sum_{i=1}^{k}\frac{T_{i.}^2}{n_i}\right) - \frac{T_{..}^2}{N} \\
\mathrm{SS}_{Error} &= \sum_{i=1}^{k}(n_i-1)s_i^2
\end{aligned}$$

Actually, we only need to compute any two of these, and we can then get the third using the fact that $\mathrm{SS}_{Total} = \mathrm{SS}_{Treat} + \mathrm{SS}_{Error}$.

The ANOVA table

It is customary (for books, scholarly journals, and computer software) to display the results of an ANOVA in the form of a table, as follows:

Source of variation	SS	df	MS	F
Treatments	SS_{Treat}	$k-1$	$\frac{\text{SS}_{Treat}}{k-1}$	$\frac{\text{MS}_{Treat}}{\text{MS}_{Error}}$
Error	SS_{Error}	$N-k$	$\frac{\text{SS}_{Error}}{N-k}$	
Total	SS_{Total}	$N-1$		

A mean square (MS) is simply a sum of squares (SS) divided by the corresponding degrees of freedom.

The ANOVA: A real example

Now let's return to the isoflavone compound example.

	Promensil	Rimostil	Placebo
n_i	84	83	85
$\overline{X}_i$	5.1	5.4	5.0
$T_{i.}$	428	448	425
s_i	4.1059	4.6023	3.1997

We get

$$\begin{aligned} \text{SS}_{Treat} &= \frac{(428)^2}{84} + \frac{(448)^2}{83} + \frac{(425)^2}{85} - \frac{(1301)^2}{252} = 7.21 \\ \text{SS}_{Error} &= (83)(4.1059)^2 + (82)(4.6023)^2 + (84)(3.1997)^2 = 3996.1 \end{aligned}$$

The ANOVA: A real example

The ANOVA table is

Source of variation	SS	df	MS	F
Treatments		2		
Error	3996.1		16.049	
Total		251		

Since the computed F statistic is not in the right tail, we do not reject H_0 at the .05 level of significance (or at any other typical level of significance).

Conclusion: The mean number of hot flashes per day is not significantly different for the three treatments.

Mean separation: Introduction

In the isoflavone compound example, we found no significant differences among the three treatment means, so the analysis is finished at that point. In other examples, however, the F test statistic will be larger than the critical value (at the chosen level of significance) and we will reject H_0. What then?

In that case we must try to discover which population means are significantly different from the rest. The textbook describes two methods for this, Bonferroni t tests and the Student-Newman-Keuls test. You may ignore these; we will use a different, and much easier, method called the "protected F method" that performs just as well.

Mean separation: Protected F method

The protected F method consists of t tests at significance level α for the equality of each pair of population means, but these are only performed if the F test at significance level α rejects the overall equality-of-means hypothesis.

In this context, the t test statistic for comparing μ_i and μ_j is

$$t_{ij} = \frac{\overline{X}_{i.} - \overline{X}_{j.}}{\sqrt{\text{MS}_{Error}\left(\frac{1}{n_i} + \frac{1}{n_j}\right)}}.$$

The appropriate critical values, taking α to be identical to that used for the overall F test, are $\pm t_{1-\alpha/2,N-k}$. We conclude that μ_i and μ_j are significantly different if t_{ij} is more extreme than either of the two critical values.

Mean separation: Example

An experiment was performed to study the psychological effects of exercise on male college students. Four groups of college men were studied:

- Exercisers (E): participants in a prescribed semester-length exercise program
- Quitters (Q): people who volunteered to participate in the exercise program but did not follow through
- Joggers (J): non-participants who, however, jog regularly
- Slackers (S): other non-participants

At the beginning and end of the experiment, a psychological test was taken by each person. The scoring on the exam was measured in such a way that a greater degree of satisfaction/confidence/happiness at the end of the experiment corresponded to a greater difference in the two exam scores taken by each person. We therefore think of the μ_i's as representing mean satisfaction levels.

At the 0.10 level of significance, we want to test $H_0 : \mu_E = \mu_Q = \mu_J = \mu_S$ versus H_a: at least one mean is different from the others.

Mean separation: Example

Results:

Group	n_i	$\overline{X}_{i.}$	s_i
E	5	57.40	10.46
Q	10	51.90	6.42
J	10	58.20	9.49
S	11	49.73	6.27

$\text{MS}_{Treat} = 158.88, \qquad \text{MS}_{Error} = 62.88$

Test statistic: $F = 2.53$

Critical value: $F_{.90,3,32} \doteq F_{.90,3,30} = 2.28.$

So we reject H_0, concluding that mean satisfaction levels of the four groups are not all equal.

Mean separation: Example

Protected F method of mean separation:

$$t_{EQ} = \frac{57.40 - 51.90}{\sqrt{62.88\left(\frac{1}{5}+\frac{1}{10}\right)}} = 1.27, \qquad t_{EJ} = \frac{57.40 - 58.20}{\sqrt{62.88\left(\frac{1}{5}+\frac{1}{10}\right)}} = -0.18$$

$$t_{ES} = \frac{57.40 - 49.73}{\sqrt{62.88\left(\frac{1}{5}+\frac{1}{11}\right)}} = 1.79, \qquad t_{QJ} = \frac{51.90 - 58.20}{\sqrt{62.88\left(\frac{1}{10}+\frac{1}{10}\right)}} = -1.78$$

$$t_{QS} = \frac{51.90 - 49.73}{\sqrt{62.88\left(\frac{1}{10}+\frac{1}{11}\right)}} = 0.63, \qquad t_{JS} = \frac{58.20 - 49.73}{\sqrt{62.88\left(\frac{1}{10}+\frac{1}{11}\right)}} = 2.44$$

Mean separation: Example

Critical values are $\pm t_{.95,32} = \pm 1.694$. So we conclude that 3 of the pairwise comparisons are statistically significant at the .10 level:

- Exercisers versus Slackers
- Quitters versus Joggers
- Joggers versus Slackers

In other words, faithful participants in the exercise program and joggers have significantly higher satisfaction levels than slackers; and likewise for joggers compared to people who begin an exercise program but quit.

Nonparametric ANOVA (Kruskal-Wallis test: Introduction

Recall that one of the assumptions for the ANOVA to be valid is that the variable of interest is normally distributed in each of the populations under study, or if not, then the sample sizes are large enough for the CLT to apply. If neither of these holds, then there is an alternative, nonparametric testing approach called the Kruskal-Wallis test.

The Kruskal-Wallis test is essentially an ANOVA of the ranks of the data (rather than of their numerical values). It extends the Wilcoxon rank sum test to k groups in the same way that the ANOVA F test extends the two-sample t test to k groups.

Kruskal-Wallis test: The hypotheses tested

The objective of a Kruskal-Wallis test is to test the null hypothesis

$$H_0 : M_1 = M_2 = \cdots = M_k$$

against the alternative hypothesis

$$H_a : \text{at least one } M_i \text{ is different from the others}$$

in such a way that the overall Type I error probability can be pre-specified.

Kruskal-Wallis test: Underlying assumptions

1. The samples are independent random samples from their respective populations.

2. The population distributions are all the same shape; they differ (possibly) only insofar as their medians are concerned.

Sketch of population distributions satisfying the last assumption:

Kruskal-Wallis test: Test statistic

As with the Wilcoxon rank sum statistic, we begin to compute the Kruskal-Wallis statistic by ranking all observations without regard to which sample they come from.

Notation and terminology:

- R_i: sum of the ranks associated with the ith sample
- $\frac{R_i}{n_i}$: sample mean rank for the ith sample
- $\frac{N+1}{2}$: grand mean rank

Kruskal-Wallis test: Test statistic

Then the test statistic is

$$H = \frac{12}{N(N+1)} \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{N+1}{2} \right)^2.$$

If H_0 is true, then we would expect each $\frac{R_i}{n_i}$ to be fairly close to $\frac{N+1}{2}$; so if H is large it casts doubt on H_0.

Computational formula for test statistic:

$$H = \frac{12}{N(N+1)} \left(\sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(N+1).$$

Kruskal-Wallis test: Critical values and P values

Rather than providing a table of critical values specific to the distribution of H under H_0, the textbook recommends obtaining critical values and P values from the chi-square distribution.

Specifically, use χ^2_{k-1}. So we reject H_0 at level of significance α if and only if

$$H > \chi^2_{1-\alpha,k-1}.$$

If we don't reject H_0, the analysis stops. If we do reject H_0, we must follow up with pairwise comparisons of medians, using a Wilcoxon rank sum test for each pairwise comparison. This is called the "protected Kruskal-Wallis method."

Kruskal-Wallis test: Example

(Problem 8.16 from textbook.) To compare the efficacy of three insect repellants, 19 volunteers applied fixed amounts of repellant to their hand and arm and then placed them in a chamber with several hundred hungry female *Culex erraticus* mosquitoes. The repellants were citronella, N,N-diethyl-meta-toluamide (DEET) 20%, and Avon Skin So Soft hand lotion. The data recorded below are the times in minutes until first bite; ranks are given in parentheses.

	Citronella	DEET 20%	Avon Skin So Soft
	5 (1)	12 (10)	6 (2.5)
	6 (2.5)	16 (14)	7 (4)
	8 (5.5)	25 (16)	9 (7)
	8 (5.5)	27 (17)	10 (8)
	14 (12)	28 (18)	11 (9)
	15 (13)	31 (19)	13 (11)
			17 (15)
R_i	39.5	94	56.5

Kruskal-Wallis test: Example

Test statistic:

$$H = \frac{12}{19(20)}\left(\frac{39.5^2}{6} + \frac{94^2}{6} + \frac{56.5^2}{7}\right) - 3(20) = 9.13.$$

Critical value: $\chi^2_{.95,2} = 5.99$.

So we reject H_0. There is statistically significant evidence that the median time to first bite for at least one repellant is different from the others.

Pairwise Wilcoxon rank sum tests for each pairwise comparison of medians show that median times to first bite are significantly different for citronella and DEET 20%, and also for Avon Skin So Soft and DEET 20%; but median times to first bite are not significantly different for citronella and Avon Skin So Soft.

Hypothesis testing for the probabilities of a distribution of a categorical variable

In all the hypothesis testing situations we've considered so far, the variable of interest was continuous, or discrete with many levels, e.g.:

- hind tibia lengths of periodical cicadas
- lifespan of male CETP variant carriers
- norephinephrine concentrations in medulla of toluene-exposed rats
- average number of hot flashes per day in menopausal women

Hypothesis testing for the probabilities of a distribution of a categorical variable

Now we consider situations where the variable of interest is categorical. Examples:

- presence of disease (present or absent)
- color morph of a salamander species (black, red-striped, or red-spotted)

In such situations, the parameter(s) of interest are the actual probabilities of each category of the variable, e.g. the probability that a randomly selected individual from the population of interest has the disease of interest.

Equivalently, we're interested in *population proportions*.

Hypothesis testing for the probabilities of a distribution of a categorical variable

In the *dichotomous* case, the categorical variable of interest has only two levels, which we generically label as "success" and "failure." Write

$p =$ probability of success $=$ proportion of successes in population.

We may want to test hypotheses about p, for example

$$H_0 : p = 0.9 \qquad \text{versus} \qquad H_a : p \neq 0.9.$$

More generally, we may want to test hypotheses about all the proportions, for example whether the ratio of three color morphs in a salamander population is Black:Red-striped:Red-spotted = 1:2:1.

The binomial and proportions tests: Introduction

Consider a situation with a single population. When the variable of interest is dichotomous, there are just two population proportions of interest: p and $1-p$. So there's just one population parameter of interest: p.

Hypotheses about p are of 3 types:

- H_0: $p = p_0$ versus $H_a : p \neq p_0$
- H_0: $p \leq p_0$ versus $H_a : p > p_0$
- H_0: $p \geq p_0$ versus $H_a : p < p_0$

Note: we've already dealt with a confidence interval for p (p. 165).

The binomial and proportions tests: Introduction

Example: Some years ago, public health officials in Atlanta decided that if less than 90% of Atlanta children under 6 had received the DPT vaccine, then they would carry out an immunization program. Here, p represents the proportion of children under 6 in Atlanta that had received the DPT vaccine, and the hypotheses of interest are

$$H_0 : p \geq 0.90 \qquad \text{versus} \qquad H_a : p < 0.90.$$

The statistical hypothesis test that can address this exists in 2 versions, depending on the size of the random sample drawn from the population:

- **the binomial test**, if $np_0 \leq 5$ or $n(1-p_0) \leq 5$
- **the proportions test**, if $np_0 > 5$ and $n(1-p_0) > 5$.

The binomial test: Test statistic

The test statistic for the binomial test is simply the number of successes in the sample, i.e.

$$S = \#\text{ of successes in sample.}$$

This will be an integer between 0 and n (sound familiar?). If H_0 is true, then we would expect S to lie near the middle of the bin(n, p_0) distribution; if H_0 is false then S is likely to lie closer to one of the extremes (0 or n).

You might have noticed that this is similar to the sign test statistic; in fact it's equivalent to the sign test when $p_0 = 0.5$.

The binomial test: P values

P value testing approach:

- If H_a is $H_a : p < p_0$, reject H_0 if $P(\text{bin}(n, p_0) \leq S) < \alpha$
- If H_a is $H_a : p > p_0$, reject H_0 if $P(\text{bin}(n, p_0) \geq S) < \alpha$
- If H_a is 2-sided, reject H_0 if

  $$P[\text{bin}(n, p_0) \leq \min(S, n-S)] + P[\text{bin}(n, p_0) \geq \max(S, n-S)] < \alpha$$

The binomial test: Beer bottling example

Beer drinkers and brewmeisters have long known that exposure to light can cause a "skunky" taste and smell in beer. In fact, chemical studies have shown how the light-sensitive compounds in hops called isohumulones degrade forming free radicals that bond to sulfur to cause the skunky taste. Most bottled beer is sold in green or brown bottles to prevent this. Miller Genuine Draft (MGD) is claimed to be made from chemically altered hops that don't break down into free radicals in light and, therefore, the beer can be sold in less expensive clear bottles. The company thinks the extra cost of a dark bottle will pay off for them only if more than 60% of beer drinkers would prefer MGD unexposed to light. In a taste test of MGD stored for 6 months in light-tight containers or exposed to light, a panel of 20 tasters preferred the light-tight beer 16 times. Should the company use dark bottles?

Note: Here $np_0 = 20(0.6) = 12$ and $n(1-p_0) = 20(0.4) = 8$, so the proportions test could be used, but the binomial test is doable and will give a more accurate answer since the sample size is rather small.

The binomial test: Beer bottling example (continued)

Hypotheses of interest: $H_0 : p \leq 0.6$ versus $H_a : p > 0.6$.

Test statistic: $S = 16$

P value:

$$\begin{aligned} P(\text{bin}(20, 0.6) \geq 16) &= 1 - P(\text{bin}(20, 0.6) \leq 15) \\ &= 1 - .9490 \\ &= .051. \end{aligned}$$

The company would probably want to do additional taste-panel testing, but if forced to make a decision based merely on this result, they would be well-advised to use dark bottles (even though the P value was a bit over .05).

The proportions test: Test statistic and critical values

When the sample sizes are such that $np_0 > 5$ and $n(1-p_0) > 5$, we can still do the binomial test, but alternatively we can get a good approximation to it using the normal approximation to the binomial. The test statistic is

$$z = \frac{S - np_0}{\sqrt{np_0(1-p_0)}},$$

or equivalently,

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$$

where $\hat{p}$ is the sample proportion of successes, S/n (defined previously on page 158 of these notes).

Critical values are either $z_{1-\alpha}$, z_α, or $\pm z_{1-\alpha/2}$ depending on whether H_a points to the right, to the left, or is two-sided.

The proportions test: Atlanta immunization example

Recall the Atlanta immunization example introduced on page 308. In order to test the hypotheses

$$H_0 : p \geq 0.9 \qquad \text{versus} \qquad H_a : p < 0.9,$$

a random sample of 537 Atlanta children under age 6 was taken. Of these, 460 had been immunized for DPT. Should the city carry out the immunization campaign?

Sample proportion: $\hat{p} = 460/537 = 0.857$.

Test statistic: $z = \frac{0.857-0.90}{\sqrt{0.90(1-0.90)/537}} = -3.32$.

Critical value (at .01 level of significance): $z_{.01} = -2.33$.

So we reject H_0. On this basis, Atlanta officials determined that it was wise to spend the \$'s to carry out the immunization campaign.

Comparing two population proportions

Just as there are situations where we want to compare the means, μ_1 and μ_2, of two populations (for a continuous variable of interest), there are situations where we want to compare the proportions, p_1 and p_2, of two populations (for a dichotomous characteristic of interest). The hypotheses to be tested may be of 3 types:

- H_0: $p_1 = p_2$ versus $H_a : p_1 \neq p_2$
- H_0: $p_1 \leq p_2$ versus $H_a : p_1 > p_2$
- H_0: $p_1 \geq p_2$ versus $H_a : p_1 < p_2$

Comparing two population proportions: Test statistic

Suppose that independent random samples of sizes n_1 and n_2 are taken from the two populations. We will consider only a large-sample version of the appropriate test (analogous to the proportions test rather than the binomial test). For this test to be applicable, we require

$$n_1 p_1 > 5, \quad n_1(1-p_1) > 5, \quad n_2 p_2 > 5, \quad n_2(1-p_2) > 5.$$

Calculate the 2 sample proportions of successes:

$$\hat{p}_1 = \frac{\text{\# successes in 1st sample}}{n_1}, \quad \hat{p}_2 = \frac{\text{\# successes in 2nd sample}}{n_2}$$

and a "pooled" sample proportion,

$$\hat{p}_c = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2} = \frac{\text{total \# of successes}}{\text{total sample size}}.$$

Rationale for pooled sample proportion: If the null hypothesis is true, then $\hat{p}_1$ and $\hat{p}_2$ are estimating the same quantity, so we get a better estimate by combining them (analogous to pooling the sample variance for the t test comparing means when we are willing to assume the two population variances are equal).

Test statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Comparing two population proportions: Critical values

This test, like the proportions test for the proportion of a single population, is based on the normal approximation to the binomial distribution. So we get critical values (and P values) from the standard normal distribution.

Critical values are $z_{1-\alpha}$, z_{α}, or $\pm z_{1-\alpha/2}$ depending on whether H_a points to the right, to the left, or is 2-sided.

Comparing two population proportions: Chronic wasting disease example

On page 166 of these notes, we described a study in which 272 deer were legally killed by hunters in the Mount Horeb area of SW Wisconsin in 2001-02. From tissue sample analysis, it was determined that 9 of the deer had chronic wasting disease (a disease similar to mad cow disease). If 272 deer from the population in that same region were sampled next winter, and 16 tested positive for the disease, would that be statistically significant evidence of a change in the infection rate?

Let p_1 and p_2 represent the proportions of infected deer in these populations in 2001-02 and 2014-2015, respectively.

Hypotheses tested: $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$

Comparing two population proportions: Chronic wasting disease example

Sample proportions:

$$\hat{p}_1 = \frac{9}{272} = .03309, \quad \hat{p}_2 = \frac{16}{272} = .05882, \quad \hat{p}_c = \frac{9+16}{272+272} = .04596$$

Test statistic:

$$z = \frac{.03309 - .05882}{\sqrt{.04596(1-.04596)\left(\frac{1}{272}+\frac{1}{272}\right)}} = -1.43$$

Critical values (taking $\alpha = .05$): $\pm z_{.975} = \pm 1.96$

So we don't reject H_0. If the sample infection rate changed to this degree, it would not constitute statistically significant evidence for a change in the population's infection rate.

Confidence interval for the difference in two population proportions

An approximate $100(1-\alpha)\%$ confidence interval for $p_1 - p_2$ can also be based on the normal approximation to the binomial distribution. The interval is

$$(\hat{p}_1 - \hat{p}_2) \pm z_{1-\alpha/2}\sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

For the approximation to be sufficiently good, the sample sizes should satisfy the conditions listed on page 321 of these notes.

Chi-square test for goodness-of-fit: Introduction

In some situations with categorical data, there are more than two categories (levels) and hence more than one proportion parameter of interest. Consider the following example.

The nests of the wood ant, *Formica rufa*, are constructed from small twigs and wood chips. As part of a study of where ants build these nests, the direction of greatest slope was recorded for 42 such nests in Pound Wood, Essex, England. Compass directions were divided into four classes: North, East, South, West. The direction of greatest slope for the 42 nests were 3, 8, 24, and 7 (in the same order of listing as the compass directions). Do the ants prefer any particular direction of exposure over another?

Chi-square test for goodness-of-fit: Introduction

This scientific question can be addressed by letting p_N, p_E, p_S, p_W represent the proportions of nests facing these directions for the entire population of wood ants (in this region), and then testing

$$H_0: \quad p_N = p_E = p_S = p_W = \tfrac{1}{4} \quad \text{versus}$$
$$H_a: \quad \text{At least one proportion is different from the others}$$

We might think we could accomplish this by doing 4 tests of population proportions equalling $\frac{1}{4}$, or 6 tests of two population proportions being equal. But neither approach allows us to prespecify the Type I error probability. So we need a new approach.

The new approach is called the chi-square goodness-of-fit test.

Chi-square test for goodness-of-fit: Hypotheses tested

For general use, let $p_1, p_2, \ldots, p_k$ represent the proportions for the k categories of the categorical variable. We aim to test

H_0 : $p_1 = p_1^0, p_2 = p_2^0, \ldots, p_k = p_k^0$ versus

H_a : At least one proportion is different from the others

Here $p_1^0, p_2^0, \ldots, p_k^0$ are proportions we specify. (In the wood ant example they are all $\frac{1}{4}$).

Note: The textbook describes H_0 and H_a in an equivalent way, but using words; it doesn't use the notation p_1, p_1^0, etc.

Chi-square test for goodness-of-fit: Test statistic

Suppose we take a random sample of size n from the population of interest, and observe the categorical variable of interest for each one. Define:

- $O_i =$ observed frequency of ith category
- $E_i = np_i^0 =$ expected frequency of ith category under H_0

Our test statistic is

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}.$$

Chi-square test for goodness-of-fit: Test statistic

Note:

- $\chi^2 = 0$ only if the observed frequencies exactly match those predicted by the null hypothesis
- The larger the discrepancy between the observed frequencies and what they are expected to be under H_0, the larger the value of χ^2.
- Therefore, we should reject H_0 for values of χ^2 that are too large.

Chi-square test for goodness-of-fit: Critical value

The test statistic is given the symbol χ^2 because it has a chi-square distribution when H_0 is true; and so we get our critical value from a chi-square distribution.

Specifically, we reject H_0 at significance level α if and only if

$$\chi^2 > \chi^2_{1-\alpha,k-1}.$$

Note: What we have just tested is what the textbook would call an *extrinsic* model. The text also describes an *intrinsic* model and modifies the χ^2 test slightly for such a model; you can skip this.

Chi-square test for goodness-of-fit: Wood ant example

Recall the wood ant example, for which the hypotheses to be tested are

H_0 : $p_N = p_E = p_S = p_W = \frac{1}{4}$ versus

H_a : At least one proportion is different from the others.

Summary of data and partial computation of test statistic:

Direction	O_i	E_i	$\frac{(O_i - E_i)^2}{E_i}$
North	3	10.5	5.357
East	8	10.5	0.595
South	24	10.5	17.357
West	7	10.5	1.167

Chi-square test for goodness-of-fit: Wood ant example

Test statistic: $\chi^2 = 5.357 + 0.595 + 17.357 + 1.167 = 24.476$.

Critical value (using $\alpha = .01$): $\chi^2_{.99,3} = 11.3$.

So we reject H_0, concluding that ants take direction into account when choosing where to build their nest. In fact, it appears (though we have not shown it statistically) that ants prefer their nests to face one particular direction (South) over the others.

Chi-square test for $r \times k$ contingency tables

A **contingency table** is a rectangular array (a table having rows and columns) of frequencies of categorical variables. The frequencies in the table can arise in either of 2 ways:

1. Taking independent random samples from r populations and recording the frequencies of one categorical variable for each sample

2. Taking a random sample from 1 population and recording the cross-classified frequencies of two categorical variables

Contingency tables example: The CASH Study

The CASH (Cancer and Steroid Hormone) study was conducted during the 1980s to investigate the relationship between oral contraceptive use and three cancers (breast, endometrial, and ovarian) in U.S. women. Part of this very comprehensive study investigated whether family history of breast cancer was a risk factor for breast cancer. 4730 women having breast cancer, and 4688 women not having breast cancer, were asked how many of their first-degree relatives (mother, sister, daughter) had breast cancer. The results are displayed below:

Breast cancer?	0	1	2 or more	
No	4403	279	6	4688
Yes	4171	511	48	4730
	8574	790	54	9418

Scientific question: Does family history of breast cancer increase a woman's own risk of breast cancer?

Contingency tables example: beetles in logs

A method commonly used by ecologists to detect an association between species (possibly mutualistic, parasitic, or something else) is to take a series of observational units where the species live or forage, such as ponds or trees, and then count the number of those units in which both species are found, neither species is found, or one or the other of the species is found.

In one such study, 500 logs in a forest were sampled for the presence of two beetle species (labeled here generically as Species A and Species B). Results were as follows:

Species A?	Present	Absent	
Present	202	80	282
Absent	106	112	218
	308	192	500

Scientific question: Do these results indicate that there is a positive association between the two species?

Chi-square test for contingency tables: Proportions

A contingency table has "cells" (row $\times$ column combinations), and the frequencies in these cells can be used to estimate the proportion of individuals in the population(s) that have the relevant characteristics.

For example, in the CASH Study, define

$$
\begin{aligned}
p_{\text{No},0} &= \text{proportion of women w/o breast cancer who have} \\
&\quad \text{0 relatives w/ breast cancer,} \\
p_{\text{No},1} &= \text{proportion of women w/o breast cancer who have} \\
&\quad \text{1 relative w/ breast cancer,} \\
p_{\text{No},\geq 2} &= \text{proportion of women w/o breast cancer who have} \\
&\quad \text{two or more relatives w/ breast cancer,}
\end{aligned}
$$

with similar definitions for $p_{\text{Yes},0}$, $p_{\text{Yes},1}$, and $p_{\text{Yes},\geq 2}$. Then estimates of, for example, $p_{\text{No},0}$ and $p_{\text{Yes},1}$ are

$$\hat{p}_{\text{No},0} = \frac{4403}{4688} = 0.9392, \qquad \hat{p}_{\text{Yes},1} = \frac{511}{4730} = 0.1080.$$

In the beetles-in-logs study, a single population is sampled, and we have for example,

$$\hat{p}_{PP} = \frac{202}{500} = 0.404, \qquad \hat{p}_{AP} = \frac{106}{500} = 0.212.$$

Chi-square test for contingency tables: Hypotheses

In words, the hypotheses to be tested are

$$H_0 : \text{The row and column variables are not associated}$$

versus

$$H_a : \text{The row and column variables are associated.}$$

Another word for "not associated" is "independent" (which your textbook uses).

Chi-square test for contingency tables: Hypotheses

In the CASH example, this is equivalent to

$$H_0 : p_{\text{No},0} = p_{\text{Yes},0}, \quad p_{\text{No},1} = p_{\text{Yes},1}, \quad p_{\text{No},\geq 2} = p_{\text{Yes},\geq 2}$$

versus

$$H_a : \text{At least one of the equalities in } H_0 \text{ is false.}$$

In the beetles-in-logs example, the hypotheses are equivalent to

$$H_0 : \frac{p_{PP}}{p_{PA}} = \frac{p_{AP}}{p_{AA}} \quad \text{versus} \quad H_a : \frac{p_{PP}}{p_{PA}} \neq \frac{p_{AP}}{p_{AA}}.$$

Chi-square test for contingency tables: Test statistic

The test statistic for testing these hypotheses is once again a chi-square statistic,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{k} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

where

- $O_{ij} =$ observed frequency of ijth cell
- $E_{ij} =$ expected frequency of ijth cell under H_0
- $rk =$ number of cells

We use double subscripts because it seems natural to do so when the data are laid out in a two-way table.

Chi-square test for contingency tables: Test statistic

The way we compute an expected frequency is different in this context than it was in the goodness-of-fit testing context, however. Here,

$$E_{ij} = \frac{i\text{th row total} \times j\text{th column total}}{\text{overall total}}.$$

For example, for the beetles-in-logs example, the E_{ij}'s are:

$$E_{11} = \frac{(282)(308)}{500} = 173.712, \qquad E_{12} = \frac{(282)(192)}{500} = 108.288$$

$$E_{22} = \frac{(218)(192)}{500} = 83.712, \qquad E_{21} = \frac{(218)(308)}{500} = 134.288.$$

Chi-square test for contingency tables: Critical values

As in the chi-square goodness-of-fit test, we reject H_0 at significance level α if and only if it is too large, and "too large" means larger than the $(1-\alpha)$th percentile of a chi-square distribution.

However, the degrees of freedom are different: they are

$$(r-1)(k-1),$$

where $r = \#$ of rows and $k = \#$ of columns.

So we reject H_0 if and only if

$$\chi^2 > \chi^2_{1-\alpha,(r-1)(k-1)}.$$

Chi-square test for contingency tables: Breast cancer example

Review the breast cancer example from several pages back. The O_{ij}'s are

Breast cancer?	0	1	2 or more	
No	4403	279	6	4688
Yes	4171	511	48	4730
	8574	790	54	9418

and the corresponding E_{ij}'s are

Breast cancer?	0	1	2 or more
No	4267.9	393.2	26.9
Yes	4306.1	396.8	27.1

Chi-square test for contingency tables: Breast cancer example

$$\begin{aligned}\chi^2 &= \frac{(4403-4267.9)^2}{4267.9}+\frac{(279-393.2)^2}{393.2}+\cdots+\frac{(48-27.1)^2}{27.1}\\ &= 106.91\end{aligned}$$

The critical value for a test at significance level 0.005 is $\chi^2_{.995,2} = 10.6$, so we reject H_0. The statistical evidence is very strong that a woman's risk of getting breast cancer is associated with her family history of breast cancer. In fact, from the sample proportions we can say that a woman's risk of getting breast cancer is *increased* if she has a family history of breast cancer.

Chi-square test for contingency tables: Beetles-in-logs example

We've already given the observed and expected cell frequencies. The test statistic is

$$\begin{aligned}\chi^2 &= \frac{(202-173.712)^2}{173.712}+\frac{(80-108.288)^2}{108.288} \\ &\quad +\frac{(112-83.712)^2}{83.712}+\frac{(106-134.288)^2}{134.288} \\ &= 27.51\end{aligned}$$

The critical value for a test at significance level 0.005 is $\chi^2_{.995,1} = 7.88$, so we reject H_0. We conclude that there is an association between the species. In fact, there is a *positive* association between species: they occur together more often than we would expect if they were just distributed randomly among logs.

Chi-square test for contingency tables: Final remarks

- There is a shortcut for finding the E_{ij}'s: compute them in the usual way for cells in the first $r-1$ rows and $k-1$ columns, but get the rest by subtraction from the appropriate row totals and column totals. Thus, in the 2×2 case, only E_{11} needs to be computed in the usual way.

- For the chi-square test to be valid, the sample sizes must be sufficiently large. Specifically, what is required is that every $E_{ij} \geq 5$.

- Some authors (ours included) recommend using a continuity correction factor in the chi-square test statistic, particularly if the dimensions of the contingency table are only 2×2. You may ignore this recommendation and stick with the χ^2 statistic described in these notes.

Correlation and regression analysis: Overview

In some scientific research, the question of interest pertains to the relationship between two continuous variables. So far in the course we have not considered situations where there are two continuous variables (though we did just consider testing for an association between two categorical variables).

Some examples of such questions:

- Is there a relationship between amount of time a child listens to Mozart in the womb and their IQ at age 16?
- Is there a relationship between GPA and amount of alcohol consumed by college students?

Correlation and regression analysis: Overview

Let X and Y represent the two continuous variables. Often, it makes sense to think that one of the two variables may depend on the other; we let Y represent that variable (the dependent variable) and let X represent the other variable (the independent, or explanatory, variable).

Statistical approach to understanding the relationship, if any, between X and Y: We imagine that X and Y exist for each member of a large (possibly infinitely large) population. We take a finite random sample of size n from the population and measure X and Y on each sampled individual.

This yields data $(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)$ which we may use to make inferences about the relationship between X and Y for the population as a whole.

Correlation and regression analysis: Overview

A useful graphical summary of such data is the scatterplot, which is a plot of the (X_i, Y_i) points (with X on the horizontal axis and Y on the vertical axis).

Terms associated with scatterplots (and with relationships between X and Y): linear, nonlinear, positive, negative, perfect, imperfect, strong, weak, nonexistent.

Example scatterplots:

Correlation and regression analysis: Overview

Correlation and regression analysis: Overview

Correlation analysis seeks to describe whether an assumed linear relationship between X and Y is positive or negative, and how strong it is.

Regression analysis seeks to quantify precisely how much Y tends to change if X changes by one unit, assuming that the relationship is linear.

Note:

- Both techniques require the relationship to be linear. Methods for studying nonlinear relationships between two continuous variables are considered in more advanced courses.
- Correlation analysis is preliminary to regression analysis.

Pearson's correlation coefficient

For correlation analysis, we need a statistic that measures the sign of the relationship, and how strong that relationship is. Such a statistic is Pearson's correlation coefficient,

$$\begin{aligned} r &= \frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sqrt{[\sum_{i=1}^{n}(X_i-\overline{X})^2][\sum_{i=1}^{n}(Y_i-\overline{Y})^2]}} \\ &= \frac{\sum_{i=1}^{n}X_iY_i-\frac{1}{n}(\sum_{i=1}^{n}X_i)(\sum_{i=1}^{n}Y_i)}{\sqrt{[\sum_{i=1}^{n}X_i^2-\frac{1}{n}(\sum_{i=1}^{n}X_i)^2][\sum_{i=1}^{n}Y_i^2-\frac{1}{n}(\sum_{i=1}^{n}Y_i)^2]}} \\ &= \frac{SS_{XY}}{\sqrt{SS_X \cdot SS_Y}}. \end{aligned}$$

Pearson's correlation coefficient

Some properties of r:

- $-1 \leq r \leq 1$
- $r = 0 \Leftrightarrow$ no linear relationship exists between X and Y
- $r > 0 \Leftrightarrow$ positive linear relationship; $r = 1 \Leftrightarrow$ perfect positive linear relationship
- $r < 0 \Leftrightarrow$ negative linear relationship; $r = -1 \Leftrightarrow$ perfect negative linear relationship
- The closer r is to 1, the stronger the positive linear relationship; the closer r is to -1, the stronger the negative linear relationship

Pearson's correlation coefficient for various scatter-plots

Inference for a population correlation coefficient: Hypotheses

Imagine computing Pearson's r for the entire population of (X, Y) values; call this quantity the population correlation coefficient, and use the symbol ρ for it.

ρ is a population parameter, and r is a point estimate of it.

We often want to address the question, "Are variables X and Y significantly linearly related?" To address this we test

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_a : \rho \neq 0.$$

One-sided H_a's could be considered too, if our research question also includes the sign (positive or negative) of the linear relationship.

Inference for a population correlation coefficient: Test statistic and critical values

The appropriate test statistic is

$$t = \frac{r - 0}{\sqrt{(1 - r^2)/(n-2)}}.$$

The appropriate critical values are:

- $t_{1-\alpha,n-2}$ if H_a is $\rho > 0$
- $-t_{1-\alpha,n-2}$ if H_a is $\rho < 0$
- $\pm t_{1-\alpha/2,n-2}$ if H_a is $\rho \neq 0$

This testing approach is valid, provided that either X and Y are normally distributed or n is sufficiently large ($n \geq 25$).

Correlation analysis example: Relationship between heart disease and a fatty diet

Data from 22 countries (from 1980) are available on variables

$$Y = 100[\log(\#\text{ of deaths from heart disease per 100,000 males aged } 55-59) - 2]$$

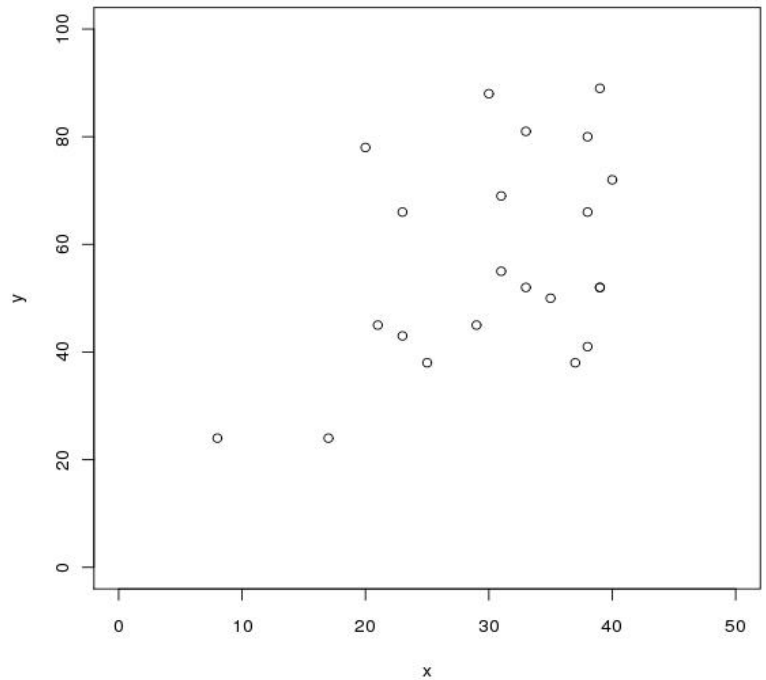
X = fat calories as a percent of total calories in diet.

Do these data indicate that heart disease and a fatty diet are associated?

Country	Y	X	Country	Y	X	Country	Y	X
Australia	81	33	France	45	29	Netherlands	38	37
Austria	55	31	Germany	50	35	New Zealand	72	40
Canada	80	38	Ireland	69	31	Norway	41	38
Sri Lanka	24	17	Israel	66	23	Portugal	38	25
Chile	78	20	Italy	45	21	Sweden	52	39
Denmark	52	39	Japan	24	8	Switzerland	52	33
Finland	88	30	Mexico	43	23	United Kingdom	66	38
						United States	89	39

Correlation analysis example: Relationship between heart disease and a fatty diet

Scatter diagram:



For these data, $r = 0.45$, indicating that there might be a statistically significant positive linear relationship.

Correlation analysis example: Relationship between heart disease and a fatty diet

Let's test

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_a : \rho \neq 0$$

at the 0.05 level of significance.

Test statistic: $t = \dfrac{0.45}{\sqrt{(1-0.45^2)/(22-2)}} = 2.23.$

Critical values: $\pm t_{.975,20} = \pm 2.086$ (P value is between 0.02 and 0.05).

So we reject H_0, concluding that there is a statistically significant linear relationship between heart disease and a fatty diet.

Another correlation analysis example: Relationship between heart disease and telephone abundance

Data from the same 22 countries from the previous example are also available on the variable

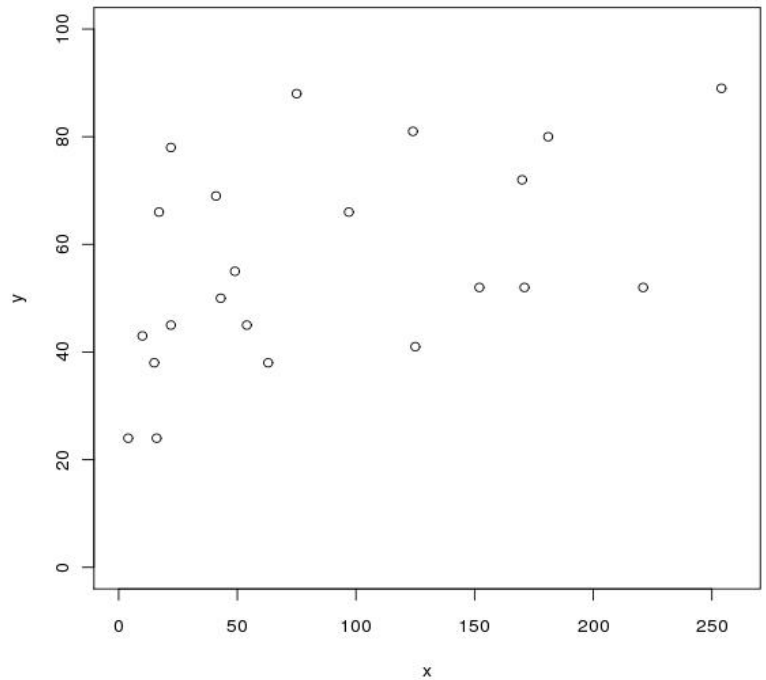
$$X = \#\text{ of telephones per 1000 population.}$$

Are heart disease (Y) and telephone abundance associated?

Country	Y	X	Country	Y	X	Country	Y	X
Australia	81	124	France	45	54	Netherlands	38	63
Austria	55	49	Germany	50	43	New Zealand	72	170
Canada	80	181	Ireland	69	41	Norway	41	125
Sri Lanka	24	4	Israel	66	17	Portugal	38	15
Chile	78	22	Italy	45	22	Sweden	52	221
Denmark	52	152	Japan	24	16	Switzerland	52	171
Finland	88	75	Mexico	43	10	United Kingdom	66	97
						United States	89	254

Another correlation analysis example: Relationship between heart disease and telephone abundance

Scatter diagram:



For these data, $r = 0.47$, indicating that there might be a statistically significant positive linear relationship.

Another correlation analysis example: Relationship between heart disease and telephone abundance

When we test

$$H_0 : \rho = 0 \qquad \text{versus} \qquad H_a : \rho \neq 0$$

at the 0.05 level of significance, we find that the test statistic is 2.37 and the critical values are the same as in the previous example ($\pm t_{.975,20} = \pm 2.086$), so we reject H_0.

The conclusion is that there is a statistically significant linear relationship between heart disease and telephone abundance. Do you think there's a cause-effect relationship here?

A proverb in science: "Correlation does not imply causation."

Nonparametric correlation coefficients

In those situations where the sample size is relatively small (< 25) and the population of values of the variable of interest is not normally distributed, we cannot safely do correlation analysis with Pearson's correlation coefficient. Instead, we use a nonparametric correlation coefficient, which is based on the ranks of the data.

Our textbook describes the following 2 nonparametric correlation coefficients:

1. Kendall's correlation coefficient
2. Spearman's correlation coefficient

Of these, Spearman's is easier to deal with so I will present it only (you're not responsible for anything regarding Kendall's).

Spearman's correlation coefficient

Spearman's correlation coefficient, r_S, is merely Pearson's correlation coefficient of the data's ranks (rather than of the original values). Specifically, we rank the X_i's from smallest to largest, and do the same for the Y_i's (settle ties as before); then plug the ranks into the formula for Pearson's r.

Two equivalent formulas for r_S (don't use 2nd one if there are ties):

1. $r_S = \frac{\sum_{i=1}^{n}(r_{Xi}-\bar{r}_X)(r_{Yi}-\bar{r}_Y)}{\sqrt{[\sum_{i=1}^{n}(r_{Xi}-\bar{r}_X)^2][\sum_{i=1}^{n}(r_{Yi}-\bar{r}_Y)^2]}}$

   where "r" is used to indicate that the observation or observations have been replaced with their ranks

2. $r_S = 1 - \frac{6\sum_{i=1}^{n} d_i^2}{n(n^2-1)}$

   where $d_i = r_{Xi} - r_{Yi}$ (the difference in the ranks of X_i and Y_i).

Spearman's correlation coefficient: Toy example

Suppose we have three (X, Y) observations, as follows:

X_i	Y_i	r_{Xi}	r_{Yi}	d_i
1.72	0.19			
0.58	0.92			
1.12	1.54			

So $r_S = 1 - \frac{6(6)}{3(8)} = -0.5$.

Spearman's correlation coefficient: Hypothesis testing

We can use Spearman's correlation coefficient to test

$$H_0 : \rho_S = 0 \qquad \text{versus} \qquad H_a : \rho_S \neq 0$$

or either one-sided version. Here ρ_S represents the value of r_S when computed using every member of the population (usually an impossible task!).

The test statistic is merely r_S itself, and the critical value(s) are found in Table C.13 in the textbook. For a specified significance level α, we use the "2-tail column" in the table if H_a is 2-sided; otherwise we use the "1-tail column." Reject H_0 if r_S is more extreme than the critical value(s).

Spearman's correlation coefficient: Examples

Let's revisit the example of heart disease versus fatty diet considered previously. For those data we have

$$n = 22, \qquad r_S = 0.39.$$

The critical values for a two-sided test at $\alpha = 0.05$ are ± 0.425, so we do not reject H_0. (In fact, $0.05 < P < 0.10$). So we conclude that there is not statistically significant evidence of a relationship between heart disease and a fatty diet.

Recall that with the parametric test (the t-test), we concluded there was statistically significant evidence of a linear relationship between these 2 variables. What gives? Two probable explanations:

- Nonparametric tests are not as powerful as parametric tests
- There are 2 outliers that strongly affect r but not r_S

Spearman's correlation coefficient: Examples

When we consider the example of heart disease versus telephone abundance, we find that

$$n = 22, \qquad r_S = 0.54.$$

The critical values for a two-sided test at $\alpha = 0.05$ are ± 0.425, so here we do reject H_0. (In fact, $0.01 < P < 0.02$). So we conclude that there is statistically significant evidence of a relationship between heart disease and a telephone abundance.

In comparison to the results based on Pearson's correlation coefficient, this analysis yields even stronger evidence against H_0. Note that there are no influential outliers in the scatterplot.

Correlation analysis: Final remarks

- Correlation analysis assumes that a linear relationship exists between Y and X (i.e. $Y_i = AX_i + B$, possibly with $A = 0$).
- Correlation analysis seeks to determine if the linear relationship is positive, negative, or 0; and how strong it is.
- Effect of transformations:
  - Linear transformations, $U_i = A_U X_i + B_U$ and $V_i = A_V Y_i + B_V$, have no effect on either r or r_S
  - Monotone increasing transformations (like taking logs) have no effect on r_S (because the ranking remains the same), but r may change
  - Non-monotonic nonlinear transformations affect r and r_S in unpredictable ways

Regression analysis: Introduction

Regression analysis adds to the results of a correlation analysis. Specifically, using a regression analysis we can address the following questions:

1. What is the equation of the straight line that best fits the data?
2. What amount of increase (or decrease) in Y can I expect by increasing X by a certain amount?
3. What do I predict the value of Y to be at a value of X that's not in my data?

Simple linear regression analysis: Introduction

Regression analysis is a HUGE topic in scientific research — many hundreds of books are devoted entirely to this subject. We will consider a very small piece of one small subtopic, called *simple linear regression analysis*.

- **Regression** refers to predicting the value of one variable from the value of others
- **Linear** refers to the assumption that the functional form of the relationship between Y (actually the mean of Y) and X is linear
- **Simple** refers to the use of merely one independent (or explanatory) variable to do the prediction of Y

Simple linear regression analysis: Conceptual foundation

We imagine that at each value of X, there is a population of values of Y. Each such population of Y-values has a mean and a variance, but this mean and variance could conceivably depend on the value of X. So we represent them by

$$\mu_{Y|X} \qquad \text{and} \qquad \sigma^2_{Y|X}.$$

Furthermore, we assume that

$$\mu_{Y|X} = \alpha + \beta X;$$

that is, we assume that $\mu_{Y|X}$ is a **linear** function of X.

Simple linear regression analysis: Conceptual foundation

Here:

- α is the "Y-intercept," while β is the "slope."
- α and β are unknown parameters, and part of our task in regression analysis is to estimate them from the available data.

How should we estimate α and β?

- By eye?
- By the method of least squares (due to Legendre in 1805)

Least squares estimation of α and β

The "least squares estimates" of α and β are those values which minimize the sum of squares of vertical deviations from the data to the line.

Picture:

Least squares estimation of α and β

Minimization methods from calculus can be used to show that the desired estimates are

$$\begin{aligned}
\hat{\beta} &= \frac{\sum_{i=1}^{n} X_i Y_i - \frac{(\sum_{i=1}^{n} X_i)(\sum_{i=1}^{n} Y_i)}{n}}{\sum_{i=1}^{n} X_i^2 - \frac{(\sum_{i=1}^{n} X_i)^2}{n}} = \frac{SS_{XY}}{SS_X}, \quad (\text{“}b\text{” in book}) \\
\hat{\alpha} &= \overline{Y} - \hat{\beta}\overline{X}. \qquad (\text{“}a\text{” in book})
\end{aligned}$$

Notice that we must obtain $\hat{\beta}$ first, then $\hat{\alpha}$.

Once we have $\hat{\alpha}$ and $\hat{\beta}$, we can plot the **least squares line**,

$$Y = \hat{\alpha} + \hat{\beta} X,$$

on the scatterplot.

Least squares estimation of α and β: Toy example

Suppose we have five (X, Y) observations, as follows:

X_i	Y_i
1	0
3	3
4	2
2	1
2	2

$\hat{\beta} =$

$\hat{\alpha} =$

Prediction using the least squares regression line

One of our main objectives in a regression analysis is to predict what the value of Y would be at any X value. Using the least squares regression line we can do this, and we can also predict what the mean of all Y-values at any X value would be.

Let x be the specified value of X at which we want to predict Y. To predict the population mean of Y-values at x, we merely plug x into the equation of the least squares regression line:

$$\hat{\mu}_{Y|x} = \hat{\alpha} + \hat{\beta}x.$$

To predict a single observation of Y at x, we do the same, i.e.

$$\hat{Y}|x = \hat{\alpha} + \hat{\beta}x.$$

Prediction using the least squares regression line

Illustrations, using the previous toy example:

Caution: Do not extrapolate too far from the interval of observed X's!

Simple linear regression analysis: Assumptions for inference

$\hat{\alpha}$, $\hat{\beta}$, and $\hat{\mu}_{Y|x}$ are all point estimates of their respective parameters. To obtain confidence intervals for these parameters and do hypothesis tests, we need to make some further assumptions. We assume that:

- $\mu_{Y|X} = \alpha + \beta X$ (as before)
- $\sigma^2_{Y|X}$ actually does not depend on X, so we write it more simply as σ^2_Y
- Each population of Y values is normally distributed
- All observations of Y are sampled independently

Simple linear regression model

All of these assumptions can be summarized and written as a model for our data, as follows:

$$Y_i = \alpha + \beta X_i + \varepsilon_i, \qquad \varepsilon_i \sim \text{independent } \mathrm{N}(0, \sigma_Y^2), \quad i = 1, \ldots, n.$$

Picture:

This model is called the **simple linear regression model**. Its parameters are α, β, and σ_Y^2.

Estimation of residual variance

We've already estimated α and β; it remains to estimate σ_Y^2. The estimate is

$$s_Y^2 = \frac{\sum_{i=1}^{n} [Y_i - (\hat{\alpha} + \hat{\beta} X_i)]^2}{n-2}.$$

This estimate is very nearly the average squared vertical deviation from the data to the least squares line; if the divisor were n rather than $n-2$ it would be exactly the average squared vertical deviation from the data to the least squares line.

We use s_Y to perform inference for many other quantities in regression analysis.

Estimation of residual variance: Toy example

For the toy example on page 378 of these notes, recall that

$$\hat{\beta} = \qquad\qquad , \qquad \hat{\alpha} =$$

Can add columns to table of data to facilitate calculation of s_Y^2:

X_i	Y_i	$\hat{\alpha}+\hat{\beta}X_i$	$Y_i-(\hat{\alpha}+\hat{\beta}X_i)$
1	0		
3	3		
4	2		
2	1		
2	2		

So $s_Y^2 =$

Confidence interval for $Y|x$

A $100(1-\alpha)\%$ confidence interval for $Y|x$ is

$$(\hat{\alpha}+\hat{\beta}x) \pm t_{1-\alpha/2,n-2} s_Y \sqrt{1+\frac{1}{n}+\frac{(x-\overline{X})^2}{SS_X}}.$$

Toy example: Confidence interval for $Y|X=2.5$ is