Introduction to Statistics

Stat 1040

Section 006

Dr. Jürgen Symanzik

Fall, 2002

Utah State University

Thanks are due to Dr. Adele Cutler and Dr. Michael Minnotte for providing us with their material (lecture notes, examples, solutions, etc.) from previous Stat 1040 courses. Thanks are also due to Dr. Natascha Vukasinovic for her valuable help in combining and enhancing the original sets of notes to their current form.

These enhanced notes have first been used in Spring 2002 for Stat 1040, Sections 002, 003, and 004, taught by Dr. Jürgen Symanzik and Dr. Natascha Vukasinovic. Please note that there have been significant updates since Spring 2002 such that the old edition of the notes should not be used instead of this new edition.

Ch. 1: Controlled Experiments

Suppose we have a new drug. How do we test whether or not it helps people in the way we wish?

We must do a _____ of the new treatment to nothing or the current treatment.

We give the new drug to subjects in the _____.

We also have a _____ that does *not* receive the new drug.

We should assign subjects to the two groups at _____.

We should do this _____, so that neither the subjects nor their doctors know which group each is in.

Ex: The Salk Polio Vaccine

In 1954, the Public Health Service organized an experiment to test the Salk polio vaccine on children in grades 1–3.

- 2 million children involved
- about 1/2 million received the vaccine
- about 1 million were left unvaccinated as part of the experiment
- the remaining 1/2 million were left unvaccinated because their parents refused to consent that they be included in the study

- Why could they get away with different size groups?

- Why not give all children the vaccine?

- Why not compare to the previous year?

- Why not give the vaccine to all children whose parents allowed them to participate, and leave the non-consenting children unvaccinated?

If differences *other* than the treatment exist between the treatment and control groups, the effects of the treatment can be _____ with effects of those other factors.

The result is known as a _____ (which might be in favor of or against the treatment).

Ex: The National Foundation for Infantile Paralysis (NFIP) proposed vaccinating all second-graders whose parents consented, leaving grades 1 and 3 as control groups. The results, in cases per 100,000 were as follows:

	Size	Rate
Grade 2 (vaccine)	225,000	25
Grades 1 and 3 (control)	725,000	54
Grade 2 (no consent)	125,000	44

What were some problems with this study?

How to assign the children to the two groups? Should the experimenters have tried to balance the two groups with respect to variables such as family income, child's health, etc.?

It is better to select the subjects randomly. With enough subjects, things will be balanced better than humans can. This is called a

_____.

- Some people will respond to the *idea* of treatment, rather than the treatment itself.
- To remove this effect, control subjects were given a _____ (in this case, an injection of saltwater).
- No subject was told which group he or she was in.
- To eliminate bias in diagnosis, the children's doctors were also not told which group each child was in.
- Since both the subjects and their diagnosticians were *blind* to which group each child was in, this is called a _____ experiment.
- A randomized controlled double-blind experiment is the best kind of experiment that can be conducted.

The randomized controlled double-blind experiment found the following rates of polio (per 100,000):

	Size	Rate
Treatment	200,000	28
Control	200,000	71
No Consent	350,000	46

Was the vaccine effective in reducing the incidence of polio?

Compare this result with the result from the NFIP study – what can we conclude?

A randomized controlled double-blind experiment minimizes bias, and can make calculations about the experiment easier.

Historical Controls

Some doctors use *historical controls*, in which the treatment group is compared to past patients treated the old way. But the groups may be non-comparable.

For example, coronary bypass surgery studies had the following three-year survival rates:

	Historical	Randomized
Surgery	90.9%	87.6%
Control	71.1%	83.2%

Can you indicate any possible bias of the historical studies?

Would you recommend coronary bypass surgery?

Ch. 2: Observational Studies

In a _____, the investigators decide who goes in the treatment and control groups (preferably randomly).

In an _____, subjects assign themselves to groups and the investigators can only watch.

Ex: Smoking studies are not controlled experiments because the experimenter cannot influence whether the subject smokes or does not smoke.

Does smoking cause lung cancer? Although there is a strong _____ between smoking and lung cancer, this does not necessarily imply _____.

There could be a _____ related to both smoking and rate of lung cancer.

A solution is to _____ possible confounding factors by comparing similar subgroups of the treatment group and control groups.

Ex: Compare male smokers age 55–59 to male non–smokers age 55–59.

Compare ...

Question:
Why is this a reasonable solution?

Ex: It has been observed throughout the year that in weeks where little ice cream has been sold, a large number of flu cases has been observed. On the other hand, in weeks where a lot of ice cream has been sold, only a few flu cases have been observed. Can we conclude that ice cream prevents flu?

Association does not imply causation!

Ex: In 1973, U.C. Berkeley looked at sex bias in their graduate programs. Out of 8,442 male applicants, 44% were admitted. Of the 4,321 female applicants, only 35% were admitted.

This looks like strong evidence of discrimination, until the individual majors are investigated. For the six largest majors, the data looks like this:

	Men		Women	
Major	Number of applicants	Percent admitted	Number of applicants	Percent admitted
A	825	62	108	82
B	560	63	25	68
C	325	37	593	34
D	417	33	375	35
E	191	28	393	24
F	373	6	341	7
Total	2,691	44	1,835	30

Now is there strong evidence of sexual discrimination?

Warning:

When evaluating studies, ask:

Have they controlled for all (reasonable) confounding factors?

Who says so?

Some Examples:

A smoking study financed by the American Cancer Society is likely to be more reliable than one financed by the Tobacco Institute.

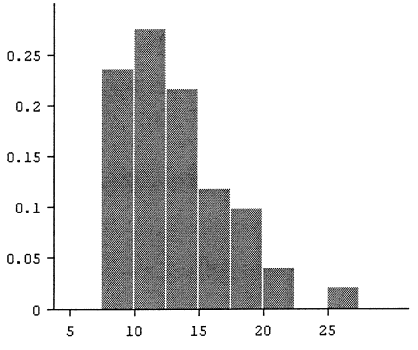
A recent study indicating that bald men are more likely to suffer heart attacks is suspicious, simply because it was financed by the Upjohn Co., manufacturer of Rogaine.

Ch. 3: Histograms

If we have a large amount of data that we wish to understand, a good start is to use the information to draw a simple plot or graph.

This idea is often called _____.

When we are interested in the distribution of a single set of numbers, we use a _____.



Ex: The following data set shows the age of people watching a recent Disney movie in a local cinema (there was no age restriction for children).

Viewers Age in Years						
13	12	11	19	24	2	13
17	15	2	17	15	7	15
13	27	4	16	13	9	5
8	19	4	17	12	5	28
7	23	13	13	6	21	20
10	6	10	7	17	18	19
10	2	13	9	27	17	14
21	9	19	12	3	18	11
18	11	25	11	10	12	14
17	5	14	30	7	15	4
19	18	11	19	1	13	8
15	20	4	4	14	13	10
15	24	14	11	22	15	7
23	15	12	18	16	6	23
12	14	23	18	10	25	18
24						

What can we conclude from these numbers (without any further work)?

We can do better by constructing a _____.

We break our set of possible values into _____ and count how many data values fall into each.

Viewers Age Distribution Table

Years	Count	Percent
0-4	10	9
5-9	16	15
10-14	33	31
15-19	29	27
20-24	12	11
25-29	5	5
30-34	1	1
	106	99

This is better, but it's still hard to grasp the entire situation at once.

We can do better still by displaying the data graphically in a *histogram*.

The area of each block is proportional to the number of data points in that class interval.

- When just looking at the histogram (and not at the distribution table), we would observe that about 25% of the viewers ages were from 15 to 19 years.

- Which percentage of viewers is in the 20 to 24 years age group?

- And which percentage in the 0 to 9 years age group?

In this example, the heights in the histogram are directly proportional to percentage, because the class intervals all have equal width. This might sometimes not be the case!

Constructing a Histogram

1. Start with a *distribution table*.

2. Set up the *horizontal axis*, making sure spacing is consistent.

3. *Calculate heights*. Since

   $$\text{percentage} = \text{area} = \text{width} \times \text{height},$$

   we must calculate the height of each block as

   $$\text{height} = \text{percentage}/\text{width}.$$

4. *Draw the blocks* using the heights just calculated.

Ex: The 1990 Census found the following breakdown of years of education for adults over 25 in Cache County:

Years of Education				
at least	but less than	Percentage	Width	Height
0	9	3		
9	12	8		
12	13	25		
13	16	34		
16	17	19		
17	23	11		

Histograms that use percentages on the vertical axis are drawn using the _____.

In the density scale, area represents percentage, height represents crowding, i.e., the percentage per horizontal unit.

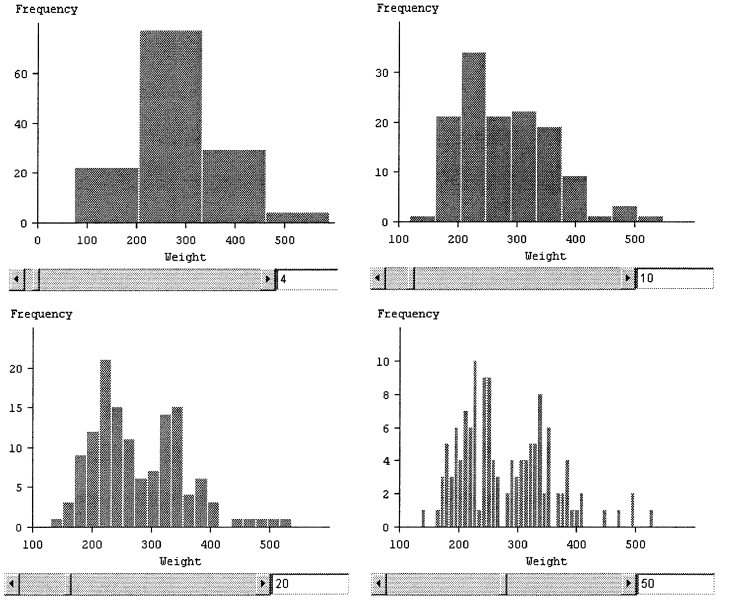
About what percentage of the over-25 population of Cache County in 1990 had at least 12 but less than 15 years of education?

With the density scale, areas of blocks are percentages. The area under the histogram over an interval equals the percentage of cases in that interval. The total area is 100%.

Warning:

The visual impression we get from a histogram is highly depending on the width of each class and the starting point of each class. Even worse, histograms with different class widths could easily provide a misleading visual impression!

Following are 4 histograms of the same data set, the weights in pounds of 132 professional male athletes. What can you conclude about the distribution of the weights when looking at each of the 4 histograms?



Now imagine what could happen if we also modify the width of individual classes, allowing classes with different widths.

Recommendation: When constructing histograms (and distribution tables), use *equal class widths* whenever possible!

Variables

A _____ is a characteristic of a person or thing which is of interest in a study (or an experiment), e.g.,

- age
- height
- weight
- income
- family size
- occupation
- race

A _____ variable is indicated by a number.

A _____ variable is indicated by categories.

Quantitative variables can be _____ or _____.

A discrete variable can only take a limited number of values. A continuous variable can take an unlimited number of values.

Histograms are only used for quantitative variables.

Ch 4: The Average and the Standard Deviation

Histograms give a good summary of the data, but sometimes they are inconvenient.

We can summarize a data set by reporting numbers that indicate the _____ and _____ of the data. These numbers are especially useful when the histogram has only one peak and is (at least roughly) symmetric.

The Average

The most commonly used measure of the center of a set of numbers is the _____, also called _____.

The average of a list of numbers equals their sum, divided by how many numbers there are in the list.

Ex: What is the average of the following list of numbers: 3, 6, 9, 7?

Ex: The Health and Nutrition Examination Survey (HANES) surveyed a cross-section of 20,322 Americans, and found that the men had an average height of 5' 9", and an average weight of 171 lbs. The women had an average height of 5' 3.5", and an average weight of 146 lbs.

By reducing data sets to their averages, we can compare many groups simultaneously.

The 2 plots (see Figure 3, page 59, in your textbook) show the height and weight data for both genders for six different age groups.

Does the height plot mean people are shrinking?

Intuitively, what percentage of people do you think has a weight below average?

And what percentage do you think has a weight above average?

The histogram (see Figure 4, page 62, in your textbook) shows a histogram for the weights of the women in the HANES sample.

In the histogram of women's weights, only about _____ of the area is to the right of the average, rather than the _____ we might expect.

To see why, consider the averages of the following sets of numbers:

- 1, 2, 2, 3; average = _____
  values below average: ___
  values equal to average: ___
  values above average: ___

- 1, 2, 2, 5; average = _____
  values below average: ___
  values equal to average: ___
  values above average: ___

- 1, 2, 2, 7; average = _____
  values below average: ___
  values equal to average: ___
  values above average: ___

A histogram balances when supported at the average.

The Median

The _____ of a list of numbers is the middle number when they are arranged from smallest to largest. If there are two middle numbers, the median is the average of these two middle numbers.

How to find the median of a data set:

1. Sort the values from smallest to largest.

2. If you have an odd number of values, the median is the center one.

   If you have an even number of values, the median is the average of the two center values.

Ex:

List 1: 3, 5, 1, 8, 0

Sorted list:

Median:

Ex:

List 2: 2, -1, 5, 1

Sorted list:

Center values:

Median:

For a histogram, half the area is to the left of the median, and half of the area is to the right of the median.

If the histogram of your data is symmetric, the median and the average will be close.

If the histogram has a long right tail, the average will be greater than the median.

If the histogram has a long left tail, the average will be less than the median.

Ex: What do you think is larger – the average or the median body weight of the women in the HANES sample? Look at Figure 4, page 62, in your textbook.

If your histogram has an extremely long tail, the mean will be strongly influenced by the few cases in the tail, and the median will better indicate the center of your data.

Ex: Why might we prefer the median to the average as a measure of the center of a list of incomes of employees of Microsoft?

The Standard Deviation

The _____ (SD) is a measure of how spread out data values are around the average.

Most numbers from a data set will be within _____ of their average.
Few will be more than _____ away from their average.
Almost none will be more than _____ away from their average.

More precisely:

- about _____ of values will be within 1 SD of the average.
- about _____ of values will be within 2 SD's of the average.
- about _____ of values will be within 3 SD's of the average.

How to calculate the SD:

1. Find the average.
2. Find the *deviations from the average* (entry - average).
3. Square the deviations from the average.
4. Average the squared deviations from the average.
5. Take the square root.

$$\text{SD} = \sqrt{\text{average of [(deviations from avg)}^2\text{]}}.$$

Ex: Find the SD of the list: 5, 12, 15, 20.

In practice, we usually use computers or calculators with statistical functions to calculate the SD. (But don't if you're asked to "show your work.")

Be aware, there's another number, slightly larger than the SD, also sometimes referred to as the "standard deviation" (we'll call it SD^+).

Most calculators use the symbol σ for the SD, and s for the SD^+.

Read Section 4.7 in the textbook to learn how to find out whether your calculator calculates SD or SD^+.

Note: in calculating the SD, we are using the *root-mean-square* operation.

$$\text{r.m.s.(list)} = \sqrt{\text{average of (entries}^2\text{)}}$$

Ch 5: The Normal Approximation to Data

The _____ or _____
is the most important curve in statistics.

This curve is that important that it is even honored on a German bank note**!**

Many data sets approximate the normal curve closely, and many statistical procedures use the normal curve.

The equation of the normal curve is

$$y = \frac{100\%}{\sqrt{2\pi}} e^{-x^2/2},$$

where $e = 2.71828...$ — but we won't use the equation.

The graph of the normal curve

- is symmetric,
- has a total area of 100% between the curve and the horizontal (x-) axis,
- and the curve is always above the x-axis (though it gets very close to the x-axis).

Many histograms fit the normal curve closely, if they are drawn in _____.

Standard units say how many SD's above (+) or below (-) the average a value is.

Ex: In the HANES sample, women 18-74 had an average height of 63.5", and an SD of 2.5".

If a woman is 66" tall, she is 66"-63.5" = 2.5" above average, which is 1 SD above average.

Therefore, 66" = +1 standard unit.

- What is a height of 56" in standard units?
- 67.5"?
- 63.5"?
- What height is -.6 standard units?

If a histogram follows the normal curve, the area under the histogram is approximately the same as the area under the normal curve.

Recall from Chapter 4: about 68% of values are often within 1 SD of the average, about 95% are within 2 SD's, and about 99.7% are within 3 SD's. This is based on the normal curve.

The Normal Table

The table on page A-105 in the textbook gives percentages within certain numbers of standard units ("z") of the average. The modified table we use in class does not contain the "Height" column.

We can calculate the area for any range of values by using this table.

Sketches often help in using this table.

Note: different books contain different tables.

Ex: What is the area under the normal curve?

- Between -0.5 and 0.5?

- Between 0 and 1.5?

- Between -2 and 0?

- Between -2 and 1.5?

- Between -1.5 and 2?

- More than 1.5?

- Less than 0.5?

- Between 0.5 and 1.5?

The Normal Approximation

We use the _____ to estimate areas for histograms which closely follow the normal curve.

How to do it:

1. Convert the original values in standard units.
2. Use the normal table to find the area under the curve.

Ex: Women in HANES: avg = 63.5", SD = 2.5"

- What percentage were between 62" and 68.5"?

- What percentage were less than 67.5"?

Percentiles

If data is not normal, then the average and SD are not as good as summary statistics.

Ex: For the 1987 income data, avg = \$44,500, and SD = \$32,000. By the normal approximation, \$0 is at -1.4 standard units.

How did we get this value of -1.4?

The area to the left of -1.4 under the normal curve is about 8%. Thus, by the normal approximation, about 8% of families should have a negative income!

With non-normal data, we can summarize more accurately than with the average and SD by using _____.

The _____ is the value such that p% of the values are below, and (100-p)% of the values are above.

We commonly use percentiles including 1, 10, 25, 50, 75, 90, and 99.

The 50th percentile is called the _____.
The 25th percentile is called the _____.
The 75th percentile is called the _____.

Percentiles of 1992 US family income

Percentile	Income
1	$1,300
10	$10,200
25	$20,100
50	$36,800
75	$58,100
90	$85,000
99	$151,800

Interquartile Range

The _____ (IQR) is

$$IQR = \text{75th percentile} - \text{25th percentile}.$$

The IQR is used as a measure of spread when the SD is too heavily influenced by one or two extreme tails.

Ex: What is the IQR of the 1992 US family income data?

Calculating Percentages

When a histogram follows the normal curve, we can use a normal table to estimate the percentiles of the data.

1. Work backwards in the table to go from area to z (if a percentile greater than 50) or -z (if less than 50).

2. Convert z or -z to the original units.

Ex: What is the 99th percentile of women's heights in the HANES study?

Ex: What is the first quartile of the women's heights??

Change of Scale

Suppose we wish to transform our data into new units (e.g., meters to feet, degrees Fahrenheit to degrees Celsius). Then we have to:

- multiply every value by the same number;
- add another number to every value.

How will this change the average and SD?

- Adding the same number to every value on a list adds the same number to the average. The SD doesn't change.
- Multiplying every value on a list by the same number will multiply the average by the same number and the SD by the absolute value of the number.
- Neither will affect the standard units.

Ex: To convert degrees Fahrenheit to degrees Celsius, we use the formula

$$C = \frac{5}{9}(F - 32) = \frac{5}{9}F - \frac{160}{9}.$$

Given Fahrenheit temperatures 32, 50, 59, 68, and 86, with average 59 and SD 18:

1. Convert the temperatures to Celsius, and find the average and the SD.

2. How could we calculate the Celsius average and SD without converting all of the values?

Ch. 6: Measurement Error

If the same thing is measured several times, several different values are likely to be obtained:

measurement =

_____ + _____

Take several measurements and average them to estimate the exact value.

The SD of all the measurements is the likely size of the chance error in a single measurement.

Ex: 100 measurements of weight of a candy bar:

ave = 10.38 grams
SD = 0.56 grams

Expect the next measurement to be around _____, give–or–take _____.

Outliers

Even careful measurements have occasional _____.

Out of 100 numbers, how many numbers would you expect to be more than 3 SD's away from the average?

Should we discard outliers in general?
This depends:

Yes, if ...

No, if ...

Warning: Do not reject valid data simply because it does not fit some theoretical curve.

Ex: As a scientist, you collect 100 measurements to check on a theory you're fond of. All but 2 of the measurements support the theory nicely, but those 2 are very different from the other 98. You're not aware of anything that went wrong when you took those 2 measurements, but you're sure something must have. Should you include those 2 values when you analyze your data?

Bias

In addition to chance error, a measuring process may also contain a _____ or _____.

This means, all measurements are systematically too high (or too low).

measurement =

_____ + _____ + _____

Ch. 8: Correlation

The methods described so far (histograms, averages, SD's, etc.) are good for looking at one variable at a time, but we need something else to look at the relationship between two variables.

Ex: Karl Pearson's 1903 study of the heights of 1,078 fathers and their (adult) sons.

A _____ or _____ allows us to study the relationship between fathers' height and sons' height (see Figure 1 on page 120 in the textbook).

In this scatterplot, each point represents a single father–son pair.

The x–coordinate of the point is the father's height.

The y–coordinate of the point is his son's height.

The variables are usually chosen so that we can think of the _____ (or _____) (x) variable as having some influence on the _____ (or _____) (y) variable.

The main mass of points is shaped roughly like a football. This is a common shape for scatterplots. (If each variable's histogram closely follows a normal curve, the joint scatterplot will be football shaped.)

The collection of points slopes up. This is called _____ (taller fathers tend to have taller sons).

If taller fathers tended to have shorter sons, there would be a downward slope to the graph, and we would call this _____.

If a son's height is always exactly equal to his father's height, all the points would fall on a straight line (_____). Knowing a father's height would tell us his son's height exactly.

If they were not exactly equal, but were always close, the points would fall in a narrow band around a straight line (_____). Knowing a father's height would give us a very good idea of his son's height.

If we had only _____, the points would fall very broadly around the line. Knowing a father's height wouldn't tell us much about his son's height.

Ex: What do you think: How much does it help to know a father's height to predict his son's height?

Ex: What type of association would you expect (positive/negative, weak/strong) between the ages of the husbands and wives in a large survey of married couples? Why?

The SD line

The line the points cluster around is called the _____.

The SD line always goes through the middle of the football, more precisely, through the point that represents the x–average and the y–average.

When the two variables have a positive association, the slope of the SD line is

slope = ...

If they have a negative association, the slope of the SD line is

slope = ...

All points on the SD line are the same number of SD's away horizontally as vertically from the respective averages.

Suppose we have a scatterplot of two variables and we want to summarize it numerically. What numbers should we use?

1. The average of x. The horizontal center of the point cloud.
2. The average of y. The vertical center of the point cloud. The *point of averages* (x-avg., y-avg.) gives the center of the cloud.
3. The SD of x. The horizontal spread of the cloud.
4. The SD of y. The vertical spread of the cloud.

These still don't describe the relationship of the two variables. We'll measure clustering around a line by

5. The _____.

The correlation coefficient is a unitless number between -1 and 1.

If r is positive, the two variables show a _____ association; if r is negative, there is a _____ association.

Numbers close to 1 or -1 show _____ and _____ correlation, respectively.

If r *equals* 1 or -1, the two variables have _____ correlation, and are falling exactly on a line.

If r is 0, we say there is _____ correlation between the two variables. This means there is no *linear* relationship between the two variables (although it is possible that there is a nonlinear one).

Weak correlations (say, .1 to .5) are common, especially in the social sciences.

Note that there is no direct way to interpret the exact value of the correlation coefficient. If r is 0.6, that doesn't mean that 60% of the values are clustered around the line, or even that it is "twice" as linear as an r of 0.3.

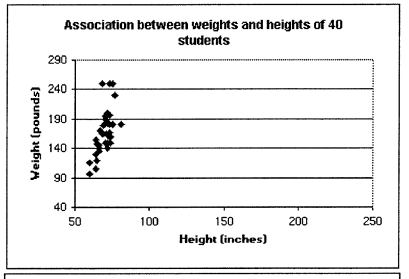
To compute r, convert each variable to standard units. The average of the products is the correlation coefficient.

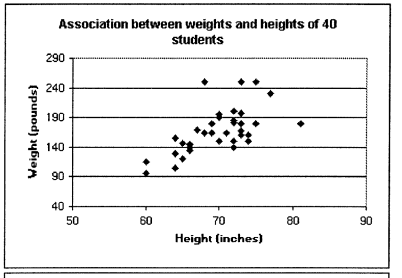
Ex:

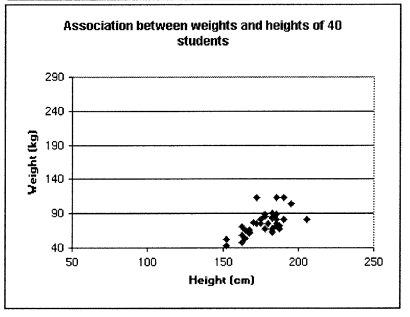
x	y	x in s.u.	y in s.u.	product
1	5			
3	9			
4	7			
5	1			
7	13			

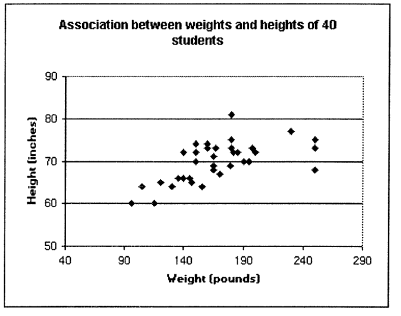
Ch. 9: More About Correlation

Question: Which plot shows the highest correlation coefficient r?









Some features of the correlation coefficient r:

- r is a pure, unitless number.
- r does not change if we
  - add the same number to each x–value
  - add the same number to each y–value
  - multiply the x–values by the same positive number
  - multiply the y–values by the same positive number
  - switch x and y

Caution

r is not always a good indicator of the relationship between x and y — remember that it only measures _____.

A generally linear pattern may have r near 0 due to _____, and a strong _____ can still lead to r near 0.

Always look at the scatterplot!

Ecological Correlations

Correlations that are based on averages or rates (and not on individual data points) are called _____.

Ecological correlations tend to be much _____ than the correlations for individuals, because ...

Note: A high ecological correlation does not imply that the association is equally strong for individual data points.

Warning:

Correlation measures _____.
But _____ is not the same as causation!

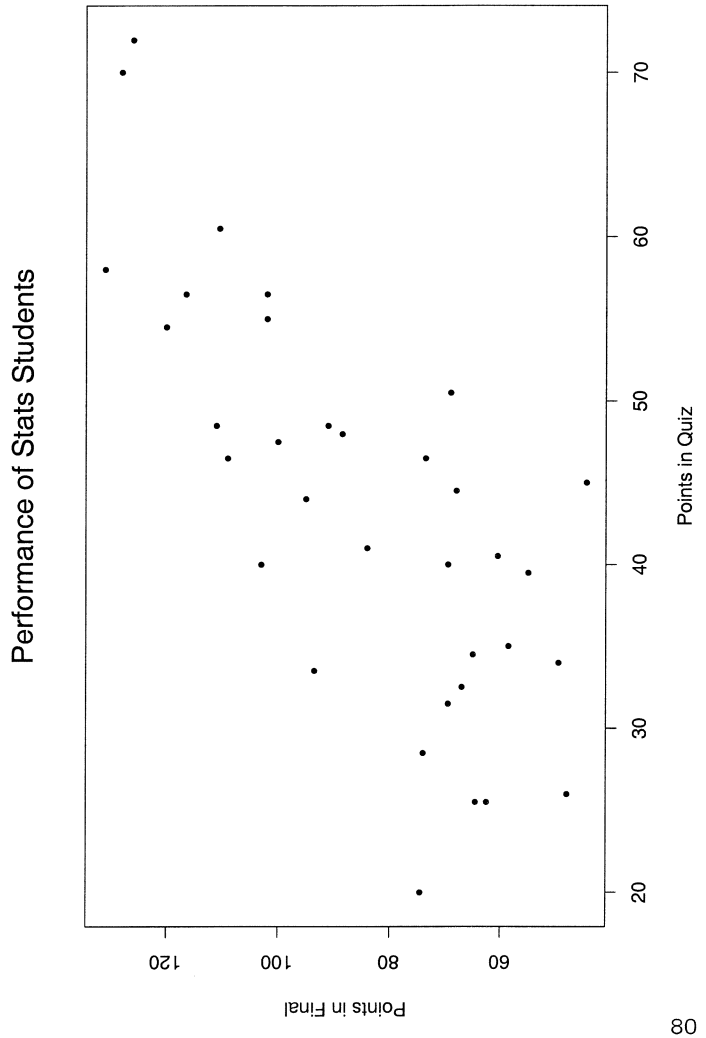
Ex: The scores on standardized reading tests of elementary school children are strongly correlated with their shoe sizes. Does improving reading skills cause foot growth?

Ch. 10: Regression

The correlation coefficient r is a number that describes the linear relationship of two variables. _____ allows us to make predictions about one variable based on a second variable.

Ex: In the "Performance of Stats Students" data, we have

quiz-avg (x) = 44 final-avg (y) = 85
$SD_{quiz} = 12.5$ $SD_{final} = 25$ r = 0.75



Question: How can we predict the points in the Final based on the points in the Quiz for a student? Can we simply use the SD line?

What else could we use?

The _____ for y on x estimates the average value of y corresponding to each value of x.

This line says that associated with an increase of one SD in x, there is only an increase of r·SD's in y.

The regression line goes through the point of averages, but its slope is ...

Ex: What is the slope of the regression line for the "Performance of Stats Students" data set?

If someone scored 56.5 points in the Quiz (i.e., one SD above avg), what is their predicted score in the Final?

If someone scored 19 points in the Quiz (i.e., two SD below avg), what is their predicted score in the Final?

In general, we can predict the y-value for any given x-value as follows:

1. Convert your independent (explanatory) variable into standard units.

2. Mulitply by r. This will give the predicted value of the dependent (response) variable in standard units.

3. Convert this predicted value back into its own units.

Ex:

If someone scored 60 points in the Quiz, what is their predicted score in the Final?

If someone scored 40 points in the Quiz, what is their predicted score in the Final?

If someone scored 0 points in the Quiz, what is their predicted score in the Final?

Question: How does the regression line look in some special cases?

- If r is 0, there is no association, and the average of y doesn't change when we change x.

- If r is 1 or -1, we have perfect association, and y changes 1 SD when x changes 1 SD.

Regression Effect

Ex: The Law School Admissions Test is a standardized test with an average score of 500 and SD 100. A large group of students took the LSAT, and then attended a LSAT-preparation class before taking it a second time. The correlation between the two scores was 0.6.

On the second test, the group of students who received a 300 on the first test improved to a 380 on average. But the students who received a 700 on the first test dropped to 620 on average. What's going on?

When correlation is less than perfect, there will be a football-shaped spread of the data points around the SD line.

This spread makes the bottom group (for the first variable) rise on average, while the top group (for the first variable) drops on average.

This is called the _____, and it is caused simply by the spread around the regression line.

The _____ is that people try to see some significance in these different outcomes.

Two Regression Lines

If there is no obvious explanatory (independent) and no obvious response (dependent) variable, then there are *two* regression lines.

Since r is symmetric, we can regress either variable on the other.

Ex: Score in Physics Quiz and Stat Quiz - which is the explanatory and which the response variable?

Ex: Age and height of children - which is the explanatory and which the response variable?

Ch. 11: The R.M.S. Error for Regression

The regression line is used to predict y from x, but the actual values won't generally fall exactly on the line.

A _____ or _____ measures how far off the actual value is from the predicted value:

residual = actual y-value – predicted y-value

Ex: (continued from Chapter 10)

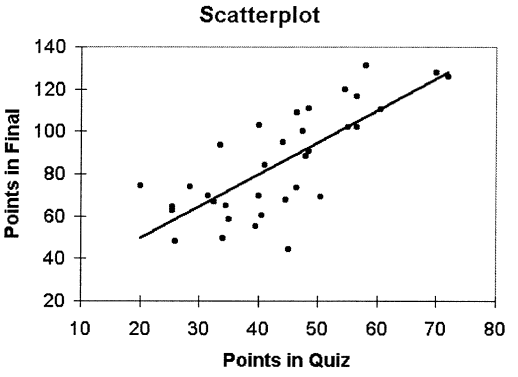
If someone scored 60 points in the Quiz, what is their predicted score in the Final?

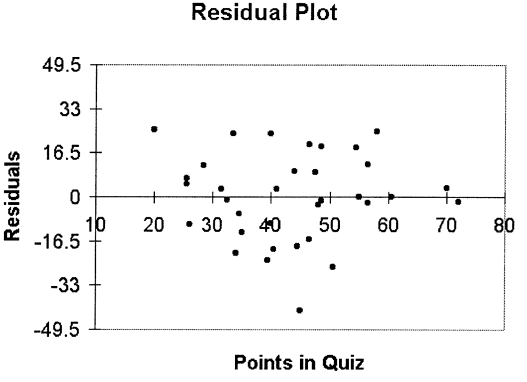
If someone scored 40 points in the Quiz, what is their predicted score in the Final?

In fact, the student with 60 points in the Quiz obtained 110 points in the Final. The residual is:

One student with 40 points in the Quiz obtained 103 points in the Final. The residual is:

Another student with 40 points in the Quiz obtained 55 points in the Final. The residual is:





We can measure the average size of our errors by using the _____:

$$\text{r.m.s. error} = \sqrt{\text{average of (residuals)}^2}$$

The *r.m.s. error of the regression line* means the same for the regression line as the SD for the average.

In many cases, about ___ % of data points will lie within 1 r.m.s. error from the regression line.

About ___ % of data points will lie within 2 r.m.s. errors from the regression line.

Computing the R.M.S. Error

We have two ways of computing the r.m.s. error of the regression line:

1. By definition:
   - Calculate the regression line.
   - Calculate the predicted values.
   - Calculate the residuals.
   - Calculate the r.m.s. of the residuals.
2. By the following shortcut:

$$\text{r.m.s. error} = \sqrt{1 - r^2} \times SD_y$$

Calculating r.m.s. errors

Ex: Predicting Final score from Quiz score:
quiz-avg (x) = 44 final-avg (y) = 85
$SD_{quiz} = 12.5$ $SD_{final} = 25$ $r = 0.75$

Ex: Predicting son's height from father's height (see Chapter 8 and page 170 in the textbook):
Fathers: $avg_x = 68''$, $SD_x = 2.7''$
Sons: $avg_y = 69''$, $SD_y = 2.7''$
$r = 0.5$

Plotting Residuals

We usually plot residuals against our explanatory (x) variable or any other known variable.

The residuals should have an average of 0 and the residual plot should show no trend.

If there is a strong pattern in the residual plot, the regression line is not appropriate.

Ch. 12: The Regression Line

Equation of the regression line:

$$y = \text{intercept} + \text{slope} \cdot x$$

where

slope =

intercept =

= predicted y when $x = 0$

Ex: In the "Performance of Stats Students" data, we have

quiz-avg (x) = 44 final-avg (y) = 85

$SD_{quiz} = 12.5$ $SD_{final} = 25$ r = 0.75

We calculate the regression line as

Using this equation, we can predict the Final scores for Quiz scores:

Quiz	Final
40	
60	

The slope says that _____ with each extra point in the Quiz, there is an increase of _____ points in the Final, on average.

We cannot fully rely on the slope to predict y when the data originates from an _____.

There are two major confounding factors here:

- _____
- _____

Remember, association is not the same as causation.

Ex: HANES men 18–24 years:

height-avg (x) = 70" weight-avg (y) = 162 lb

$SD_X = 3$" $SD_Y = 30$ lb

$r = 0.47$

Regression equation for predicting weight based on height:

Regression equation is:

Note: Sometimes the intercept does not make sense; it may be negative when we would expect it to be zero or positive.

The Method of Least-Squares

Among all possible lines one can draw on a scatterplot, the regression line has the smallest possible sum of the squared residuals, i.e., the smallest r.m.s. error. For this reason, the regression line is also called "least squares" line.

Ch. 13: What are the Chances?

The terms "chance," "the chances," or "probability" are used loosely all the time.

Probability was initially developed to solve gambling problems.

For our purposes, we need a more rigorous definition.

_____ is the most common interpretation of probability. The chance of something equals the percentage of the time it is expected to happen if the basic process is repeated over and over again, independenly and under the same conditions.

The simplest examples are games of chance, such as those dealing with coins, dice or cards.

Play a Game of Chance

Each person in class flips a coin once:

Number of Heads (H's):

Number of Tails (T's):

Total:

Proportion of H's:

Play again!

What proportion of H's do you expect when we flip the coins 1,000,000 times?

Chance (or probability) makes predictions about long-run behavior (of a coin, die, or whatever).

Chance (or probabilty) = Proportion in the long run.

Ex: Toss a coin – the chance of H is _____.

Ex: Roll a die – the probability of a "6" is _____.

Facts on Chances

Chances are always between 0% and 100%.

Something impossible has a 0% chance of occurring. Something certain happens 100% of the time. Everything else is in between.

Ex: Chance of 7 in a die roll:

Ex: Chance of a number between 1 and 6 in a die roll:

_____ are the same as chances, but are usually expressed as a decimal or fraction, rather than a percentage.

0% chance	=	probability 0
100% chance	=	probability 1
40% chance	=	probability 0.4
	etc.	

_____: The chance of something equals 100% minus the chance of the opposite thing.

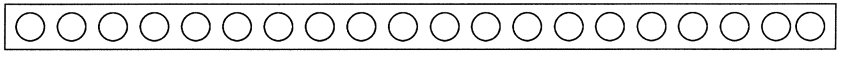
Ex: If you play a game, and the chance of winning is 45%, the chance of not winning is _____.

Ex: If we roll a die, we have a probability of $\frac{1}{6}$ to roll a 6. The probability to roll a 1–5 (that is, to *not* roll a 6) is _____.

Drawing Balls from a Box

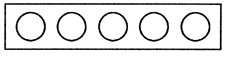
Ex: Suppose we draw a ball at random out of a box, and win $1 if the ball is black. We win nothing if the ball is white. If the box contains 4 black balls and 6 white ones (Game A), what is our chance of winning?

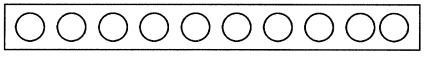
Ex: If the box has 5 black and 15 white (Game B), what is our chance of winning?



Which box is better: Game A or Game B?

Ex: And which game would you prefer in this scenario?

Game C: 

Game D: 

Note:

The probability of drawing a black ball $=$

$$\frac{\text{number of black balls}}{\text{total number of balls}}$$

Drawing Tickets from a Box

e.g., $\boxed{1}\boxed{2}\boxed{3}\boxed{4}\boxed{5}\boxed{6}$

When drawing more than one ticket, we can draw _____ or _____.

If the draws are made _____ replacement, we draw a ticket, record its number, and then put the ticket back. The chance of getting a certain number doesn't change from draw to draw.

If the draws are made _____ replacement, we draw a ticket, record its number, but then keep the ticket out. The chance of getting a certain number changes in each draw, based on the results of the previous draw.

Ex: Two draws with replacement from

$\boxed{1}\boxed{2}\boxed{3}$

If the first draw is a $\boxed{2}$, then the second ticket is drawn from the box

$\boxed{}$.

The chance of drawing a $\boxed{1}$ as the second ticket is:

Ex: Two draws without replacement from

$\boxed{1}\boxed{2}\boxed{3}$

If the first draw is a $\boxed{2}$, then the second ticket is drawn from the box

$\boxed{}$.

The chance of drawing a $\boxed{1}$ as the second ticket is:

Playing Cards

Spades: ♠	Clubs: ♣	Diamonds: ♢	Hearts: ♡
Ace	Ace	Ace	Ace
2	2	2	2
3	3	3	3
4	4	4	4
5	5	5	5
6	6	6	6
7	7	7	7
8	8	8	8
9	9	9	9
10	10	10	10
J	J	J	J
Q	Q	Q	Q
K	K	K	K

Ex: Shuffle the pack of 52 cards and deal one.

What is the probability that this card is a ♡?

What is the probability that this card is the Queen of hearts?

What is the probability that this card is a red card?

Conditional Probabilities

Ex: Now consider shuffling the 52 cards and drawing two, placing them face down.

What is the probability that the second card is the Queen of hearts?

Suppose we know the first card is the King of Spades. Now, what is the probability that the second card is the Queen of hearts?

The first probability (or chance) is called _____ – we know nothing.

The second probability (or chance) is called _____ – we know what the first card was.

_____: The chance that two things will happen equals the chance of the first, multiplied by the chance of the second given that the first has happened.

Ex: Two cards are dealt from a well-shuffled deck. What is the chance that both are spades?

Ex: A coin is tossed twice. What is the chance of a head followed by a tail?

Independence

Two things are said to be _____ if the chances for the second given the first are the same, no matter how the first turns out. Otherwise, they are *dependent*.

Ex: Are coin tosses independent?

Ex: If we draw with replacement from (1, 1, 2, 2), are the two draws independent?

Ex: What if we draw from the same box *without* replacement?

Ex: If we roll a die twice, are the 2 rolls independent?

Note:
Random draws with replacement are independent.

Random draws without replacement are dependent.

The Multiplication Rule for Independent Events

If two things are independent, the probability that both will happen equals the product of their unconditional probabilities.

Note that this is a special case of the general Multiplication Rule (discussed earlier in this chapter).

Ex: Draw 2 cards with replacement from the pack of 52 cards. What is the chance of a ♡, followed by a Queen?

Ex: A couple has 2 children. We assume that the gender of one child is independent from the gender of its siblings.

What is the probability that both children are girls?

And what is the probability that one child is a boy and the other child is a girl?

Ch. 14: More About Chance

Ex: The roll of a die can give a value of 1, 2, 3, 4, 5, or 6.

The chance of getting a value from {4, 5, 6} is:

The chance of getting a value from {1, 5} is:

We must be careful that all of the ways have equal probability.

Ex: If we roll and sum two dice, we get a number from 2, 3, 4, 5, 6, 7, 8, 9, 10, 11, or 12.

Is the probability we get a 7 equal to $\frac{1}{11} = 9\%$?

There are 21 different two dice combinations:

1-1, 1-2, 1-3, 1-4, 1-5, 1-6,
2-2, 2-3, 2-4, 2-5, 2-6,
3-3, 3-4, 3-5, 3-6,
4-4, 4-5, 4-6,
5-5, 5-6, and
6-6.

Should the chance of a total of 7 be $\frac{3}{21} = 14\%$?

You can figure out chances by listing all the ways an event can happen.

Gambling with 3 Dice

17th century Italian gamblers bet on the sum of 3 dice. They reasoned:

The chance of getting 9 is the same as the chance of getting 10 because there are 6 ways of getting 9 and 6 ways of getting 10:

Ways of getting 9:

1-2-6, 2-3-4,
1-3-5, 2-2-5,
1-4-4, 3-3-3

Ways of getting 10:

1-4-5, 1-3-6,
2-2-6, 2-3-5,
2-4-4, 3-3-4

However, it turned out that 10 came up slightly more often than 9.

They asked Galileo about this, who reasoned as follows: Suppose we have a white die, a grey die, and a black die. We have $6 \times 6 \times 6 = 6^3 = 216$ different combinations.

There is only 1 way to get the numbers 3,3,3:

white = 3, grey = 3, black = 3

There are 3 ways to get the numbers 3,3,4:

white = 3, grey = 3, black = 4
white = 3, grey = 4, black = 3
white = 4, grey = 3, black = 3

There are 6 ways to get the numbers 1,2,6:

white = 1, grey = 2, black = 6
white = 1, grey = 6, black = 2
white = 2, grey = 1, black = 6
white = 2, grey = 6, black = 1
white = 6, grey = 1, black = 2
white = 6, grey = 2, black = 1

Mutually Exclusive Events

Two events are _____ if both cannot happen together.

For each pair of events, are they mutually exclusive?

Ex: The first card is a heart, the first card is a spade?

Ex: The first card is a heart, the second card is a spade?

Ex: The first card is a heart, the first card is an ace?

Ex: The white die is a 1, the black die is a 1?

Ex: The white die is a 1, the sum of white and black dice is 12?

Ex: Two independent events?

_____: If two things are mutually exclusive, the chance that (at least) one of them will happen is the sum of their individual chances.

Ex: When we roll a die once, the chance of a 1 *or* a 2 is:

Ex: When we draw a card from a well-shuffled deck, the chance of a ♡ *or* a ♢ is:

Ex: When we roll two dice, the chance of getting at least one 6 is:

Ex: When we draw a card from a well-shuffled deck, the chance of a ♡ *or* a Queen is:

What if we want at least one of things which aren't mutually exclusive?

Possibility 1: If *two* events are *not* mutually exclusive, the chance at least one will happen is

$$\text{chance(1st)} + \text{chance(2nd)} - \text{chance(both)}.$$

Ex: When we roll two dice, the chance of getting at least one 6 is:

Ex: When we draw a card from a well-shuffled deck, the chance of a ♡ *or* a Queen is:

Note that this formula only works for *two* events!

Possibility 2: Use the opposites rule. Calculate the chance of the opposite (which will be that *none* of the things occur), and subtract that chance from 100%.

Ex: When we roll two dice, the chance of getting at least one 6 is:

Ex: When we draw a card from a well-shuffled deck, the chance of a ♡ *or* a Queen is:

Ch. 16: The Law of Averages

John Kerrich, a South African mathematician, tossed a coin 10,000 times in a row, and kept track of the results:

Number of Tosses	Number of Heads	Difference from Half (= Chance Error)
10	4	-1
20	10	0
50	25	0
100	44	-6
200	98	-2
500	255	5
1,000	502	2
2,000	1,013	13
5,000	2,533	33
10,000	5,067	67

The Law of Averages

If we toss a coin many times

number of heads =

half the number of tosses + chance error

The _____ says that the chance error (for a large number of tosses) is likely to be

- large in absolute terms
- small compared to the number of tosses

The percentage of heads is likely to get closer to 50%, although it is not likely to be exactly 50%.

Ex: A coin will be tossed and you win \$1 if the number of heads is exactly equal to the number of tails. Which is better for you, 10 tosses or 1000?

Ex: A coin will be tossed and you win \$1 if the percentage of heads is between 40% and 60%. Which is better for you, 10 tosses or 1000?

Ex: You are betting on tosses of a coin: if the coin lands on H, you win \$ 1, if it lands on T, you lose \$1. The last 10 tosses have all been H's. What is the chance that the next toss is a H?

Box Models

Chance processes are affected by _____, e.g.,

- number of heads when tossing a coin
- amount won when playing a game of chance
- percentage of unsatisfied customers in a random sample

_____ help us answer the question:

"How big is the chance error likely to be?"

When we make a box model we must answer three questions:

- What numbers go into the box?
- How many of each number?
- How many draws will we make?

We use the term _____ as shorthand for "drawing tickets at random with replacement from a box, and then adding them up".

Ex: You play a game in which you roll a die 10 times and get paid the amount shown on the die (each time). Find a box model:

Ex: You play a game in which you roll a die 10 times. Each time a "6" occurs, you win $ 10, otherwise you lose $ 1. Find a box model:

Ex: A multiple–choice quiz has 20 questions, each with 4 possible choices. Each correct answer is worth 5 points, and for each incorrect answer you lose 2 points. Find a box model for your test score if you guess all the answers:

Summary

Setting up a box model:

- The tickets show the amounts (points, etc.) that can be won or lost on one play (question, etc.).
- The chance of drawing any particular number must be the chance of winning or losing that amount (points, etc.) on one play (question, etc.).
- The number of draws from the box equals the number of plays (questions, etc.).

Ch. 17: The Expected Value and Standard Error

If we repeat a chance process many times, we will get many different results.

In many repetitions of a chance process, the results will vary around the _____ (EV), with the amounts they are off being similar in size to the _____ (SE).

The expected value of a sum of random draws with replacement from a box equals:

$$\text{EV}_{\text{sum}} = \text{number of draws} \times \text{box average.}$$

_____:

The standard error of a sum of random draws with replacement from a box equals:

$$\text{SE}_{\text{sum}} = \sqrt{\text{number of draws}} \times \text{box SD.}$$

Ex: You play a game in which you roll a die 10 times and get paid the amount shown on the die (each time). The box model is:

What is the amount of money you *expect* to win?

What is the SE for the amount of money you win?

Ex: You play a game in which you roll a die 10 times. Each time a "6" occurs, you win $ 10, otherwise you lose $ 1. The box model is:

What is the amount of money you *expect* to win?

What is the SE for the amount of money you win?

Ex: A multiple–choice quiz has 20 questions, each with 4 possible choices. Each correct answer is worth 5 points, and for each incorrect answer you lose 2 points. The box model for your test score (if you guess all the answers) is:

What is the number of points you *expect* to get?

What is the SE for the number of points you get?

Using the Normal Curve

When the number of draws is quite large, we can use the normal curve to calculate chances associated with the sums of draws.

We convert to standard units by using the expected value (instead of the average) and the standard error (instead of the SD).

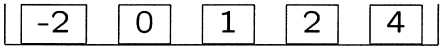
Standard units say how many SE's we are above or below the EV.

Ex: A multiple–choice quiz has 100 questions, each with 4 possible choices. Each correct answer is worth 1 points, and for each incorrect answer you get 0 points.

Find the chance that someone gets at most 30 points if he or she guesses all the answers.

Find the chance that someone gets at least 60 points if he or she guesses all the answers.

Ex: If a computer repeatedly draws 50 values from



and sums them, what percentage of the sums will be between 35 and 65?

Classifying and Counting

If you are interested in the number of times an event occurs, the box has [0] 's and [1] 's.

Ex: Suppose we are rolling a die. If we want the sum of the rolls, we use the box

1	2	3	4	5	6

On the other hand, if we just want to count 6's, we replace the tickets above with

0	0	0	0	0	1

In 60 rolls, what is the EV of the number of 6's?

And the SE?

There is a useful shortcut for calculating the average and SD of a box that contains only *two* different numbers:

$$\text{average} = \frac{(\text{smaller} \times \text{how many}) + (\text{bigger} \times \text{how many})}{\text{how many tickets in the box}}$$

$$\text{SD} = (\text{bigger - smaller}) \times \sqrt{\frac{\text{fraction}}{\text{bigger}} \times \frac{\text{fraction}}{\text{smaller}}}$$

Ex: You play a game in which you roll a die 10 times. Each time a "6" occurs, you win $ 10, otherwise you lose $ 1.

Find the average and SD of the box.

If the numbers in the box are $\boxed{0}$'s and $\boxed{1}$'s, these formula become:

$$\text{average} = \frac{\text{number of } \boxed{1} \text{ 's}}{\text{how many tickets in the box}}$$

$$\text{SD} = \sqrt{\text{fraction of } \boxed{1} \text{ 's} \times \text{fraction of } \boxed{0} \text{ 's}}$$

Ex: What is the average and the SD of the box:

0	0	0	0	0	1

Ex: 10% of people in a large population are "underweight". If we take a random sample of 200 people from this population, what is the chance that more than 21 will be "underweight"?

Ch. 18: Normal Approximation for Probability Histograms

A _____ is a histogram of chances and not of data.

For a sum of draws, each possible value for the sum has a rectangle (of width 1) centered on the value.

The area of the rectangle is the chance of that value.

If we repeat a chance process several times, we can construct an _____ of the observed outcomes.

The more repetitions we do, the more the empirical histogram will look like the probability histogram.

Ex: We toss 3 coins and count the number of heads.

What is the corresponding box model?

What is the expected number of heads?

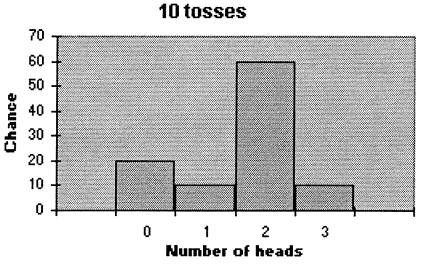
What is the standard error?

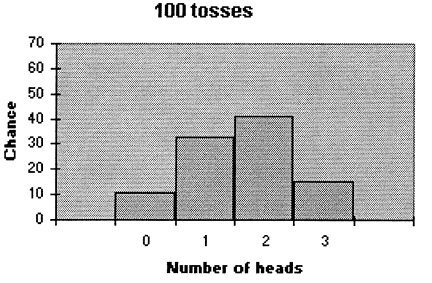
We repeat this experiment 10 times.

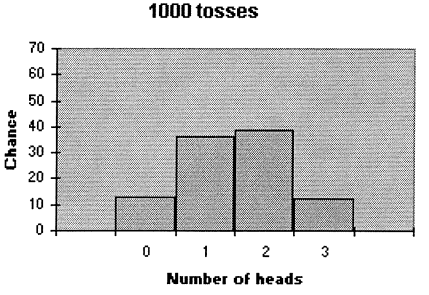
How does the empirical histogram look like?

How does the probability histogram look like?

Empirical Histograms for 10, 100, and 1000 Tosses







The Central Limit Theorem

The probability histogram of a _____ of a large number of independent draws from a box follows the normal curve even if the contents of the box do not look like a normal curve, provided that the number of draws is large.

To convert to standard units, we use EV_{sum} and SE_{sum}.

How large is "large"?

The answer depends on the box.

- If the box is symmetric, or nearly so, we can get a pretty good approximation by 30 or even less draws.

- If the box is somewhat asymmetric, the number of draws will have to be larger to get a good approximation.

- 100 draws will give a pretty good approximation to all but *extremely* asymmetric boxes.

Note:
The normal curve is associated with the _____.

Other functions of many draws (such as _____) will usually *not* follow a normal curve.

Summary

- If we draw many times from a box, the probability histogram of the sum will follow a normal curve, even if the box does not.
- The EV fixes the center of the distribution, and the SE fixes the spread. To calculate probabilities, we should put our results in standard units.
- To calculate the EV and SE requires: the number of draws, the average of the box, and the SD of the box.

Ex: Suppose we are interested in rolls of a standard, six-sided die. Are the following well represented by the normal curve?

1. The empirical histogram of 1000 rolls?
2. The probability histogram of the sum of 50 rolls?
3. The empirical histogram of 1000 sums, each of 50 rolls?
4. The probability histogram of the product of 50 rolls?
5. The empirical histogram of 1000 products, each of 50 rolls?

Ch. 19: Sample Surveys

Frequently, we like to make some statements about a class of individuals, the _____.

Ex: In predicting the results of a U.S. presidential election, the population are all the voters.

Ex: When asking about student interest in basketball at USU, the population are all USU students (and not only those attending a particular game).

It is usually impractical to measure or interview the whole population. Instead, we take a _____ from the population.

We use the sample to make _____ about the population, i.e., we generalize from the sample to the population as a whole.

We are interested in numerical values describing the population, called _____.

We estimate these parameters by _____, i.e., numbers calculated from the sample.

The sample should be _____ for the population.

Unfortunately, to check this requires knowledge of the population – exactly what we don't have!

So we must choose the sample very carefully.

The *Literary Digest* Poll of 1936

In the 1936 presidential election, Franklin Delano Roosevelt (a Democrat), was running for his second term against Alf Landon, the Republican governor of Kansas.

Literary Digest magazine did a poll and received 2.4 million responses. They predicted Landon would win 57% to 43%. On election day, however, FDR won 62% to 38%. This was the largest error ever made by a major poll.

How did the *Digest* go so wrong? Their sample was more than big enough. Unfortunately, their selection method showed a _____, where some groups were more likely to be counted than others.

The *Digest* mailed questionnaires to 10 million people from phone books, club membership lists, and magazine subscription lists.

What was wrong here?

The poll also suffered from a _____ (only 24% responded). Why was this a problem?

The Presidential Election of 1948

In 1948, three major polls all predicted Thomas Dewey would beat Harry Truman by about 5% in the presidential election. Instead, Truman won, (50% to 45%).

The polls used a _____: each interviewer must interview a certain number of subjects in each of several categories (broken down by gender, race, income, etc.), selected to match the country as a whole. But within those constraints, the interviewers can freely choose whom they interview.

This choice resulted in more Republicans being interviewed in each group.

Probability Methods

When using _____, interviewers cannot choose whom they interview. The sample is selected randomly.

The most basic example is called _____ (SRS). All sets of the proper size are equally likely to be selected.

A SRS is like putting all individuals' names in a box and drawing without replacement until we reach the desired sample size.

Ex: Ask every fifth student that enters the TSC whether they support women's basketball at USU. SRS – yes/no?

Ex: Ask all students in randomly selected classes whether they support women's basketball at USU. SRS – yes/no?

Ex: Obtain all student SSNs from the Registrar's Office and let the computer randomly select 500 of these numbers. Ask students with these SSNs whether they support women's basketball at USU. SRS – yes/no?

Even with a well-designed sampling scheme, there are many problems which make taking surveys difficult.

_____ is a problem even with face-to-face interviews. This happens more frequently with telephone surveys and mailed questionnaires.

We must always worry about the biases that the wording of the questions may bring.

Ex: Three different ways to ask about abortion:

- "The U.S. Supreme Court has ruled that a woman may go to a doctor to end pregnancy at any time during the first three months of pregnancy. Do you favor or oppose this ruling?" (47% favor, 44% oppose).

- "The U.S. Supreme Court has ruled that a woman may go to a doctor for an abortion at any time during the first three months of pregnancy. Do you favor or oppose this ruling?" (41% favor, 48% oppose).

- "As far as you yourself are concerned, would you say that you are for or against abortion, or what do you think?" (36% favor, 59% oppose).

Ex: College students watched a film of a car crash, and were asked:

- "About how fast were the cars going when they contacted each other?"
(Average answer: 31.8 mph)

- "About how fast were the cars going when they collided with each other?"
(Average answer: 40.8 mph)

We must always be aware of the possibility of respondents lying, or simply not knowing the correct answer.

Ex: In 1991, the National Survey of Men was conducted. 3,321 respondents were interviewed. These men (aged 20–39) were questioned in their homes by female interviewers. The questions concerned the men's sexual practices, and 30% refused to be interviewed. How trustworthy are such results as that the median number of sexual partners is 7.3, or that only about 1% of men are exclusively homosexual?

Ex: A political pollster once asked people what they thought of the Pepper-Johnson bill being debated by Congress. Most people expressed strong opinions for or against the bill, even though it didn't exist! Apparently, giving an uninformed opinion is easier than admitting ignorance.

Ex: _____:

Who do you think answers the "Consumer Product Survey of America?"

Are those people who answer representative for the entire US population?

Ch. 20: Chance Errors in Sampling

When the box contains $\boxed{0}$'s and $\boxed{1}$'s, the percentage of $\boxed{1}$'s in the draws will be around

$$EV_{\%} = \%\text{ of } \boxed{1}\text{ 's in the box}$$

The corresponding SE is

$$SE_{\%} = \frac{SE_{sum}}{\#draws} \times 100\%$$

i.e., $SE_{\%} = SE_{sum}$ converted to a percentage.

Ex: A certain university has 4,000 male students and 6,000 female students. If we sample 100 students without replacement, what percentage of our sample do we expect to be female? How much do we expect this to vary by?

Our sample is like drawing 100 times without replacement from the box:

The box average is:

The box SD is:

For a sample of size 100, the $EV_{\%}$ is:

The $SE_{\%}$ is:

If we sample 400 students, what will be the $EV_{\%}$ and $SE_{\%}$ of women?

Note:
Multiplying the sample size by some factor divides the $SE_{\%}$ by the square root of the factor.

Ex: In a population of 100,000 telephone subscribers, 20% earn more than \$50,000. In a random sample of 400 of these subscribers, how many do you expect to earn over \$50,000? What is the corresponding SE?

In the sample of 400, what percentage do you expect to earn over \$50,000. What is the corresponding $SE_{\%}$?

And what is the $SE_{\%}$ if the sample size is 1,600? And for 3,600?

Using the Normal Curve

Provided the sample size is large, the normal curve can be used to figure out chances.

Ex: 25% of college students own an automobile. If we take a random sample of 400 of these students, what is the chance that less than 20% own an automobile?

Accuracy of Samples

When we estimate a percentage based on a sample survey, we find:

- The sample size makes a big difference in the SE.

- Provided the sample size is small compared to the population size (10% or less), the accuracy of samples is determined by the sample size itself (and not the size relative to the population).

Ex: The Gallup Poll can use a sample of a few thousand people to estimate percentages for the entire country (200 million voters or so).

Ex: If we do simple random samples of size 2,500 of voters in every state, will the SE of the percentage of Republicans in the Utah sample be much smaller than that of California? (You may assume the SD of the "box" in each case is essentially 0.5.)

Ex: (Old Exam question):
A local politician is interested in estimating the percentage of voters who are opposed to the marriage tax. She can only afford to sample 1000 people. Other things being equal, to get equal accuracy in Logan and Salt Lake City, she would sample:

1. 500 people in Logan and 500 people in Salt Lake City.

2. more people in Salt Lake City than in Logan.

3. more people in Logan than in Salt Lake City.

Ch. 21: The Accuracy of Percentages

So far, we assumed we knew what was in the box (at least, the average and SD), e.g.:

- Games of chance

- Sampling from a known population (known average and SD)

In practice, usually, we don't know the population (i.e., the box).

That's why we sample!!

Ex: We want to estimate the percentage of voters in a large population who support the Republican candidate and obtain a SE of this percentage.

HOW?

- Estimate the population percentage using the sample percentage.

- Estimate the SE by pretending that the population is just like the sample and calculate $SE_{\%}$.

This is called _____.

Suppose we sample 1600 voters from this population, and find out that 56% of the sample support the Republican candidate.

1. Estimate the percentage of voters in the whole population who support the Republican candidate.

2. Find the standard error of this percentage.

Confidence Intervals

The normal approximation tells us that there is a 95% chance that a sample sum or percentage will be within 2 SE's of the EV.

Therefore, if we consider a range

$$\text{sample \%} \pm 2 \cdot \text{SE},$$

there is about a 95% chance that this will include the EV (the population percentage).

We call this range a _____ (CI).

For this interval to be valid, the sample size should be large and the sample percentage should not be too close to 0% or 100%.

Ex: What is the 95% confidence interval for the percentage of voters who support the Republican candidate?

Ex: Lemons are premium–grade, table–grade, or juice–grade. A farmer takes a random sample of 500 lemons from a large crop and finds that 75 are juice–grade. Find a 95% confidence interval for the percentage of juice–grade lemons in the crop.

Confidence Intervals (ctd.)

We can construct different confidence intervals. Our intervals are based on the normal approximation and the _____ we desire.

sample % $\pm$ $1 \cdot$ SE $\rightarrow$ 68% CI
sample % $\pm$ $2 \cdot$ SE $\rightarrow$ 95% CI
sample % $\pm$ $3 \cdot$ SE $\rightarrow$ 99.7% CI

Ex: The 68% CI for the percentage of voters who support the Republican candidate is:

Ex: The 99.7% CI for the percentage of juice–grade lemons in the crop is:

Confidence Intervals (ctd.)

When we want some other confidence interval, the multiplier comes from the normal curve.

E.g., for a 77% confidence interval, we use

sample % $\pm$ ____ $\cdot$ SE

Ex: A health inspector takes a random sample of 300 ten–year–olds in a city and finds that 73% of them have had chicken–pox. Find a 90% confidence interval for the percentage of 10–year–olds in the city who have had chicken–pox.

Interpretation of Confidence Intervals

The 90% confidence interval for the percentage of children who have had chicken–pox is:

Ex: True or False and explain: There is a 90% chance that the population percentage is between ___ % and ___ %.

Interpretation

We say: "We are 90% confident that the population percentage is between ___ % and ___ %".

What does this mean?

- 90% of all such intervals contain the population percentage

or

- Before you sample, there is a 90% chance the sample you get will give an interval containing the population percentage.

Ex: Suppose we have a box containing a very large number of marbles, 80% red and 20% blue. If we have 100 people each sample 2,500 marbles and construct a confidence interval as

$$\text{percent reds in sample} \quad \pm \quad 2 \cdot \text{SE},$$

then about 95 (95%) should include the true value (80%) in their intervals.

In a computer simulation of the process, 96 of 100 confidence intervals include 80%. The other 4 do not.

Ch. 23: The Accuracy of Averages

Looking at the average of a sample is similar to looking at percentages in a sample.

Ex: Suppose we roll a die 100 times. The box model is:

The average of the box is 3.5, and the SD is 1.7.

The sum of the 100 rolls will be about _____, give or take about _____.

If the sum is 350, the average will be: _____

If the sum is 1 SE low, 350 - 17 = 333, the average will be: _____

If the sum is 1 SE high, 350 + 17 = 367, the average will be: _____

The average of the rolls will be about 3.5, give or take about 0.17.

Note:

The EV of the average of draws is:

$$EV_{avg} =$$

The SE of the average of draws is:

$$SE_{avg} =$$

Normal Approximation for Averages

Recall: When drawing at random with replacement, the sum of the draws will approximately follow the normal curve for a large number of draws.

Also, the percentage of 1's in the draws (from a 0-1 box) will approximately follow the normal curve for a large number of draws.

Finally, the average of the draws will also approximately follow the normal curve for a large number of draws.

Ex: If we make 100 rolls of a fair die, what is the chance that the average of the rolls will be greater than 3.6?

And what is the chance that the average of 400 rolls is greater than 3.6?

Note:
As with percentages, increasing the number of draws by a factor will decrease the SE of the average by the square root of the factor.

The Sample Average

Usually, we don't know the box, just the sample.

As in the percentages case, we assume the box is well-represented by the sample.

Ex: A telephone company wishes to estimate the average length of weekend long-distance calls. A random sample of 50 calls gives an average length of 15 minutes, and an SD of 7 minutes.

Find the average length of all weekend long-distance calls and the corresponding SE:

CI's for the Average

We can calculate confidence intervals in the same way as for percentages – with a large sample, we can be 95% confident that the true population (box) average is within 2 SE's of the sample average.

Ex: For the telephone call example, a 95% confidence interval for the average length of a call is:

A 90% confidence interval for the average length of a call is:

What do these CIs mean?

Ex: Educational level is measured for a sample of 400 people, and the average is found to be 11.6 years, with an SD of 4.1 years. What is a 95% confidence interval for average educational level?

Is the normal approximation reasonable?

Which SE?

- SE for sum $= \sqrt{\text{number of draws}} \times$ box SD
- SE for average $= \frac{\text{SE for sums}}{\text{number of draws}}$
- SE for count = SE for a sum from a 0-1 box
- SE for percentage $= \frac{\text{SE for count}}{\text{number of draws}} \times 100\%$

- If we know the box, we reason forward. The chance quantity will be near the EV, but probably off by an SE or so.

- If we don't know the box, we reason backward. We estimate the EV by the chance quantity, but recognize that we are probably off by an SE or so, which we estimate by estimating the box SD from the sample.

Note:
The formulas for simple random samples don't apply in other cases!

Ch. 24: A Model for Measurement Error

Recall from Chapter 6:

If our measurement technique is unbiased,

individual measurement =

_____ + _____

The Gauss Model for Measurement Error

The _____ says:

- Chance error is like a draw from a box called the "_____"
- The error box has an average of zero provided the measurements are made with no bias.

SO WHAT?

The methods from Chapter 23 can be used to get confidence intervals for the exact value provided that:

- The Gauss model holds with no bias;
- The number of measurements is large.

The 95% CI is:

$$\text{AVG of sample measurements} \pm 2 \cdot SE_{AVG}$$

SE_{AVG} is calculated in the usual way, but we approximate SD_{box} with the SD of the sample measurements.

Ex: The speed of light was measured 2,500 times. The average was 299,774 kilometers per second, and the SD was 14 kilometers per second. Assume the Gauss model, with no bias. Find a 95% CI for the speed of light.

Ex: In the previous example, light was timed as it covered a certain distance. The distance was measured 57 times, and the average of these measurements was 1.594265 kilometers. True or False and explain: We can construct a 95% CI for the distance based on the information provided.

We can calculate the SE_{AVG} of any list of numbers, but that SE (and all the confidence intervals) will make sense only if the variability in the data is like the variability in repeated draws from a box.

The Gauss model does not apply if there is a trend or pattern in the data.

For example:

- Temperatures
- Stock market
- Rainfall
- Retail sales

Which SD?

Ex: A chemistry student weighs a sample 100 times and finds the average weight is 38.13 g with an SD of 0.9 g. Find a 95% confidence interval for the true weight, assuming the Gauss model with no bias and no pattern and no trend in the data.

Ex: Another student weighs a similar chemical 30 times with the same method and has an average 39.26 g and SD 1.1 g. Find a 95% CI for the true weight.

Note: The SD belongs to the measuring procedure, not the thing being measured. If there is a choice, use the most reliable SD, which will usually be the one from the largest number of measurements.

Exercises

Ex: Temperature measurements are taken at USU every hour during the month of February. Can we use our methods on these ($28 \times 24 =$) 672 measurements to calculate a confidence interval for average February temperature at USU? Why or why not?

Ex: A machine makes sticks of butter whose average weight is 4.0 oz with a SD of 0.05 oz. There is no trend or pattern in the data. There are 4 sticks in a package.

1. A package weights _____, give or take _____ or so.

2. A store buys 100 packages. Estimate the chance that they get 100 lb of butter, to within 2 oz.

Ch. 26: Tests of Significance

Ex: Jim, Jane, and Peter have been playing a game of Monopoly for several hours....

Jim (shouting):

"The dice are not fair – they are biased. There are far more double sixes than there should be!"

Jane (worried):

"It seems you have studied a little bit too much of Stat 1040 – why should those dice not be fair? We just bought the game – the dice have not been used before and nobody had a chance to manipulate them!"

Jim (not convinced at all) takes a pen and paper and starts to record the outcome of the next 720 rolls....

After several hours ....

Jim (jumps up and shouts):

"Here it is – the dice are not fair! During the last 720 rolls, there have been 32 double sixes. We all know, the chance for a double six should be _____. Thus, the expected number of double sixes after 720 rolls is _____."

Jane (a bit sceptical):

"But that could just be due to chance error, couldn't it?"

Jim (convincingly):

"Maybe, but if the dice really are fair, there is only a small chance of getting 32 or more double sixes. Peter – can you help me and calculate that exact chance?"

Peter:

"Sure the chance is: ..."

Jane (now convinced):

"Hmmmm... That's indeed a pretty small chance. We better use another set of dice when we play Monopoly next time."

Conclusion:

In the previous game of Monopoly, there is only a very small chance that we will see this many double sixes in 720 rolls. This means:

- The players have been extremely lucky, or
- The dice are really biased: The true probability of double sixes is higher than 1/36.

Tests of Significance

The previous example shows the general pattern of _____ or _____.

We want to show that some observed difference is real (i.e., the difference is in the population, not just in the sample).

- A skeptical person sees a difference and says the difference is merely due to chance variation.
- A credulous person sees the difference and says there is a real difference.
- A statistically literate person will find out who is right, i.e., how likely an observed difference is with nothing to explain it *but* chance variation.

The Null and the Alternative

In significance testing, we compare two _____:

- A _____ which says that the difference is due to chance variation.
- And an _____, which says that the difference is real. This is usually what we wish to prove.

A null hypothesis must be formulated as a box model.

Note:
Hypotheses are about the numbers in a box, not in a sample.

Null and Alternative for Monopoly

Ex: In our Monopoly example, it is:

Null Hypothesis:

Alternative Hypothesis:

Test Statistic

Once we have our null and alternative hypothesis formulated as a box model, we can calculate a _____.

A common test statistic for averages and percentages is the _____:

$$z = \frac{\text{observed - expected}}{\text{SE}}.$$

A test using the z-score is called a _____.

Note:
The test statistic z represents how many standard units an observed value is from its expected value, calculated from the null hypothesis.

z-score for Monopoly

Ex: The z-score for the Monopoly example is:

Note:
The z-score is not very meaningful itself. So we use the normal curve to convert z into an _____, commonly called _____ (for probability).

P-values

The P-value is the chance of getting a sample value at least as extreme as the observed one, given that the null hypothesis is true.

Ex: In our Monopoly example the P-value is:

Conclusion

If the P-value is small (less than 5%), there is strong evidence against the null hypothesis. This means, we _____ the null hypothesis.

If the P-value is fairly large (greater than 5%), the evidence against the null hypothesis is weak. This means, we _____ the null hypothesis.

We do not say that we _____ the null hypothesis, because we may just not have enough data to see that the null hypothesis is wrong.

Ex: In our Monopoly example, the conclusion is:

Statistical Significance

If the P-value is less than 5%, the result is _____.

If the P-value is less than 1%, the result is _____.

Note:
The P-value is not the chance that the null hypothesis is true – it is the chance of seeing such a weird sample if the null hypothesis were true.

So, if the P-value is small, we think the null hypothesis is not true.

Four-Step Procedure for Hypothesis Testing

1. State the null and the alternative hypotheses in words and in terms of a box model.

2. Find the test statistic (i.e., standard units).

3. Calculate the P-value (area under the curve).

4. State conclusions (rejecting the null hypothesis or not rejecting, and in your own words).

Ex: Bottles of orange juice are supposed to have 16 fluid ounces. A random sample of 100 bottles from a large batch contains an average of 15.7 ounces with an SD of 0.2 ounces. Test the hypothesis that the bottles are being filled correctly, against the alternative that they are not full enough.

Ex: Researchers separately asked 153 husbands and wives to state the highest school grade completed by the wife. For each couple, they recorded

X = husband's answer - wife's answer.

X averaged 0.32, with an SD of 1.1. Does this suggest an average difference which is greater than 0?

t-tests

The t-test is used instead of the z-test when:

- the number of draws is small (less than 30), and
- the SD of the box is unknown (i.e., we must bootstrap), and
- the histogram for the tickets in the box is close to the normal curve.

The t-test is just like the z-test, but:

1. We use SD_+ instead of SD, where

   $$SD_+ = \text{SD of measurements} \times \sqrt{\frac{\text{number measurements}}{\text{number measurements - 1}}}.$$

2. We still calculate our test statistic as

   $$\frac{\text{Observed - Expected}}{\text{SE}},$$

   but we do not use the normal table to find our P-value.

   Instead, we use a _____, using the line with _____ equal to the number of measurements - 1.

The Gauss Model for Measurement Error

If our measurement technique is unbiased,

individual measurement =

_____ + _____

The _____ says:

- Chance error is like a draw from a box called the "_____"

- The error box has an average of zero provided the measurements are made with no bias.

Ex: Sugar packets are supposed to have a weight of 1500 grams. A student buys 5 packets and finds:

$$1473, 1489, 1525, 1585, 1513$$

Average:

SD:

Is there more sugar in the packets than it is supposed to be? Assume the Gauss model with no bias.

When to use the t-test?

Number of observation is

Histogram of the box is

SD of the box is

Ch. 27: Two-Sample Tests

In this Chapter, we want to look at *two* independent samples from two populations.

Ex: : Suppose we have two boxes (populations): box A and box B.

Box A	Box B
$\text{AVG}_{box} = 110$	$\text{AVG}_{box} = 90$
$\text{SD}_{box} = 60$	$\text{SD}_{box} = 40$

Suppose we draw 400 from box A and 100 from box B and compare the two sample averages:

Sample from Box A	Sample from Box B
$\text{EV}_{avg} =$ ____	$\text{EV}_{avg} =$ ____
$\text{SE}_{sum} =$ ____	$\text{SE}_{sum} =$ ____
$\text{SE}_{avg} =$ ____	$\text{SE}_{avg} =$ ____

We expect the sample average for box A to be _____ higher than the sample average for box B, give or take _____ .

Ooooops – we need another SE!

The SE of the difference of two sample averages is:

$$\text{SE}_{\text{Diff}} = \sqrt{(\text{SE}_{avg_1})^2 + (\text{SE}_{avg_2})^2}.$$

Thus, for our two samples, SE_{Diff} is:

So, we expect the sample averages to differ by 20, give or take _____ .

Similarly,

$$\text{SE}_{\text{Diff}_{\%}} = \sqrt{(\text{SE}_{\%_1})^2 + (\text{SE}_{\%_2})^2}.$$

Ex: : County A has 45% Republicans. County B has 47% Republicans. We take a random sample of 300 people from County A and 500 from County B.

We expect the percentage of Republicans in the samples to differ by _____, give or take _____.

The two-sample z-test

We want to know if the two population (i.e., box) averages (sums or %) are the same.

Step 1: State hypotheses:

Null hypothesis:
Average for box A equals average for box B.

Alternative hypothesis:
Average for box A is greater / less than average for box B.

Step 2: Calculate test statistic:

$$z = \frac{\text{observed difference - expected difference}}{SE_{Diff}}.$$

or

$$z = \frac{\text{AVG of Sample a - AVG of Sample b}}{SE_{Diff}}.$$

Step 3: Obtain P-value:

Same as before - use the normal curve.

Note: The AVG of this normal curve will be zero, the SD will be SE_{Diff}.

Step 4: State conclusions:

Same as before: Based on the P-value, decide whether or not to reject the null hypothesis, and explain your findings in your own words.

Ex: 100 freshman women and 100 freshman men are each given a "Survey of Study Habits and Attitudes." Each individual receives a score from 0 (very poor study habits) to 200 (very good study habits).

The 100 women have an average score of 120, with an SD of 28. The 100 men have an average score of 105, with an SD of 35. Is this reason to believe that the population of freshman women has better study habits than that of freshman men, on average?

Ex: A large university takes a simple random sample of 200 male students, and another simple random sample of 300 females. As it turned out, 107 of the men in the sample used a personal computer on a regular basis, compared to 132 of the sample women. Is there a real difference between the percentages of men and women who use a PC on a regular basis, or is it just a chance variation?

Experiments

This two-sample testing method is commonly used in experiments.

Ex: Two hundred volunteers are used as subjects in a randomized controlled experiment on the effects of regular doses of vitamin C.

The 100 randomly assigned to the treatment (vitamin C) group averaged 2.3 colds, with an SD of 3.1 colds over the period of the experiment. The 100 members of the control group, receiving a placebo, averaged 2.6 colds, with an SD of 2.9 colds. Is there strong reason to believe that regular doses of vitamin C reduce the risk of colds?

When does the two-sample z-test apply?

We can use the two-sample z-test:

- When we have two independent random samples from two populations that we want to compare;
- For randomized controlled experiments – then we *pretend* that the treatment group and the control group are independent random samples.

We can't use the two-sample z-test:

- When our samples are not random;
- When our samples are not independent;
- When we have the whole population – then we know everything about these populations and we can compare them directly – no need for any test!

Ex: An investigator wants to show that first–born children score higher on IQ tests than second–borns. In one school district, he finds 400 two-child families with both children enrolled in elementary school. He gives these children an IQ test and obtains the following results (scores were adjusted for age differences):

– 400 first–borns: AVG = 29; SD = 10.
– 400 second–borns: AVG = 28; SD = 10.

Can he use the two–sample z–test to see whether or not his assumption is true?

Ex: Suppose I teach a class of 200 students, 100 men and 100 women. I give a comprehensive examination, and I find that the average score for the men is 75.4 with an SD of 10.2 and the average score for women is 78.5 with an SD of 10.8 (both sets of scores follow the normal curve closely). Is it appropriate to use the two-sample z–test to decide whether the men and women have different average exam scores?

Ch. 28: The Chi-Square Test

The χ^2 – test (chi-square, pronounced "ki-square") helps us answer questions such as:

- Are the data consistent with a given chance model, or are they far off?
- Has someone manipulated the data to make them fit the chance model?
- Are two things independent in the population from which the sample was drawn?

Ex: A gambler is accused of using a "loaded" die. A record has been kept of the last 60 draws:

4	3	3	1	2	3	4	6	5	6
2	4	1	3	3	5	3	4	3	4
3	3	4	5	4	5	6	4	5	1
5	4	6	3	3	3	5	3	1	4

Summarized, the results are:

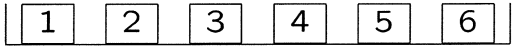
Value	Observed frequency	Expected frequency
1	4	
2	6	
3	17	
4	16	
5	8	
6	9	

Are these results consistent with what we would expect from a fair die? Or are there too many 3's and 4's?

The χ^2–test

Step 1: State hypotheses:

Null hypothesis:
The die is fair. Rolling this die is like drawing at random with replacement from the box:



Alternative hypothesis:
The die is loaded. Rolling this die is not like drawing at random from the above box.

Step 2: Calculate test statistic:

We combine all our frequency data into one value that measures how well the model is doing.

This value is called the _____.

Calculating χ^2–statistic

$$\chi^2 = \text{sum of } \frac{(\text{observed freq. - expected freq.})^2}{\text{expected frequency}}$$

In our die example:

Value	Obs.	Exp.	(Obs.– Exp.)^2	$\frac{(Obs.-Exp.)^2}{Exp.}$

$\chi^2 =$ _____

Step 3: Obtain P-value:

The P-value is obtained from the _____ with

degrees of freedom = # of categories − 1

The P-value is the area to the right of the calculated χ^2.

To find the approximate P-value, we use the _____ which is similar to the t–table.

In our die example:

d.f. = _____

P-value: _____

Step 4: State conclusions:

Note: The χ^2–test is valid only if all expected frequencies are 5 or more!

Applications in Genetics

Ex: According to a certain genetic theory, inheritance of flower color in *Mirabilis jalapa* (four o'clock) plants shows the following pattern: when a plant with red flowers is crossed with a plant with white flowers, all of the offspring have pink flowers. However, when the plants of this second generation are crossed with each other, 1/4 of the plants in the resulting third generation have red flowers, 1/2 of the plants have pink flowers, and 1/4 of the plants have white flowers.

You bought some third generation "four o'clock" seeds and got 88 plants with red flowers, 201 with pink flowers, and 111 with white flowers. Is there evidence that your seeds were "manipulated"?

χ^2–test for "Four o'clock" plants

Testing Independence

Ex: Independent random samples of workers in three parts of the country were asked whether they considered unemployment or inflation the more serious problem. We can sum up their responses in a *2×3 table* (or more generally, an *m×n table*).

	Northeast	Midwest	Southwest	Total
Unemp.	87	73	66	226
Infl.	113	77	84	274
Total	200	150	150	500

Do we have reason to believe that workers in different parts of the country feel differently about these issues?

χ^2–test for independence

Step 1: State hypotheses:

Our model looks like:

NE	??×[U] ??×[I]	200 draws
MW	??×[U] ??×[I]	150 draws
SW	??×[U] ??×[I]	150 draws.

Null hypothesis:

i.e., the percentage of [U] tickets is the same in each box.

Alternative hypothesis:

i.e., at least one box is different.

Step 2: Calculate test statistic:

$$\chi^2 = \text{sum of } \frac{(\text{observed freq. - expected freq.})^2}{\text{expected frequency}}$$

The observed frequencies are the entries in our 2×3 table.

The expected frequencies are obtained assuming independence.

45% of the workers surveyed (_____) consider unimployment to be more important, whereas 55% of the workers (_____) consider inflation to be more important.

If the null hypothesis is true, we expect that the same percentage of workers in all three regions feel the same way.

Region	Total	U (45%)	I (55 %)
NE	200	...	...
MW	150	...	...
SW	150	...	...

We usually use the short–cut formula:

Expected frequencies for an $m\times n$ table can be found for each cell by multiplying the row total by the column total and dividing by the total for the entire table.

When we finish, we get a table with both observed and expected frequencies.

Obs. | Exp.	Northeast		Midwest		Southwest	
Unemp.	87	90	73	68	66	68
Infl.	113	110	77	82	84	82

Now we can calculate the χ^2-statistic as we did before:

$\chi^2 =$ _____

d. f. $= (m-1)\times(n-1) =$ _____

Step 3: Obtain P-value:

Step 4: State conclusions:

Ex: A simple random sample of 200 Utah schoolchildren were asked whether or not they like math. 102 kids were boys, 41 of whom said they liked math. The other 98 were girls, 29 of whom said they like math. Is liking math independent of gender for Utah schoolchildren?

Ch. 29: A Closer Look at Tests of Sifnificance

Always report the P-value of a test, don't just say " the result was statistically significant".

Also:

- Summarize data
- Say which test was used.
- Report *exact* P-value whenever possible (not just P-value $< 5\%$ or P-value $< 1\%$).

Is the Result Significant?

We reject the null hypothesis if the P-value is large / small (circle one)

If the P-value is less that _____ the result is "statistically significant".

If the P-value is 4.9% we reject / do not reject null hypothesis (circle one).

If the P-value is 5.1% we reject / do not reject null hypothesis (circle one).

Is this really such a difference????

Moral:
Do not take the 5% and 1% levels too seriously!

Doing Many Tests

If the null hypothesis is true, we have a ____ % chance of rejecting it.

If we do 100 tests and all the null hypotheses are true, we expect to reject _____ of them!

Therefore, some of these null hypotheses might be rejected just by chance!

Moral:
Report *all* tests, not just significant ones!

Data Snooping

Usually, researchers decide which hypotheses to test only after they have seen the data.

This is called _____.

Data snooping makes P-values hard to interpret.

Ex: Test whether a coin is fair. In 100 tosses, we get 61 H's.

Null hypothesis:

Alternative hypothesis:
The coin is biased: Chance of H is *over* 50%.

or

The coin is biased: Chance of H is *not* 50%.

Note:

- Researchers like one–tailed tests because it is easier to get "significant results".
- It is more correct to decide which test to use before we look at the data.
- If there is any doubt, a two–tailed test should be done.

Is the result important?

Ex: Differences between rural and urban children on a vocabulary test average 1 point out of 50. The sample sizes are large, so the SE's are both very small, and this result is highly *statistically* significant, but of no practical importance.

Ex: Factory components differ in average size by day of the week (Monday, Wednesday, or Friday) but the differences were well within the tolerances for the component, and hence of no practical significance.

For a **large sample**, even tiny differences can be statistically significant, but it does not mean they are important.

For a **small sample**, even an important difference may not be statistically significant.

In the End ...

Watch out for:

- Tests where the data are the whole population (especially two–sample tests, χ^2 – tests...)

  IF YOU CAN'T MAKE A BOX MODEL, DON'T DO A SIGNIFICANCE TEST!

- Non – random samples (especially convenience samples);

- Badly designed experiments.

— The End —