

Lecture Outline for Sociology V3212: Statistics/Methods

Aaron Gullickson

December 20, 2005

[*Note: These lecture notes are not a substitute for class attendance. They are provided by the instructor as a class aid, but the reader assumes all risk of errors, typos, and the like.*]

Contents

1	Describing Data	3
1.1	Administrative and philosophical overview	3
1.2	The idea of a distribution	6
1.3	Measures of center and spread	10
1.4	Measuring relationships with categorical variables	15
1.5	Measuring relationships between quantitative variables	22
1.6	Interpreting OLS regression and transformations	27
1.7	Multivariate regression	34
1.8	A real research example	41
1.9	Regression diagnostics and cautions: outliers and influential points	42
1.10	Regression diagnostics and cautions: aggregate data, collinearity, and causality	46
1.11	Catch-up and review	50
2	Data Collection and Statistical Inference	51
2.1	Data collection (experiments and surveys)	51
2.2	Sampling distributions and probability rules	55
2.3	Random variables and their probability distributions	58
2.4	The sampling distribution	62
2.5	Sampling distribution, the binomial case	64
2.6	Confidence intervals and hypothesis tests	66
2.7	Errors in hypothesis testing: statistical significance and power	69

2.8	t-tests and the t-distribution	71
2.9	Contingency tables and Chi-squared tests	76
2.10	Inference for bivariate regression	79
2.11	Inference for multivariate regression	85
2.12	Non-parametric tests	87
2.13	Catch-up and review	89

1 Describing Data

1.1 Administrative and philosophical overview

- Administrative
 - Introduction
 - hand out syllabus, go over it
 - courseworks web site
 - Go over textbook
 - **Bring a calculator**
 - Homework
 - * due each Tuesday
 - * Show your work fully, be legible
 - * understand where you went wrong on each HW
 - Midterm (11/1) and Final (comprehensive, equations provided)
 - Office hours, early in week - come in with questions about HW already done
 - Who here is anxious about statistics? No need to be if you follow a few simple principles in your work.
 - * Don't memorize, understand - we will always focus on being able to explain in an intuitive fashion - you can always look formulae up.
 - * Keep up with work - everything we learn builds on other things - If you procrastinate you will have severe problems.
 - * When you make a mistake, fix what went wrong - correct your own homework, come to office hours - if you don't get a concept, it will hamper you later in the class.
 - Other Questions?
- Introduction
 - What are we doing? The purpose of this course is to teach you the fundamentals of statistical analysis.
 - The goal is to give you the ability to understand statistical material presented in leading sociology journals. Some of you may go on to do this kind of research, others may not - but you all need the ability to understand one another's research.

- Therefore, our emphasis will be on interpretation of results more than anything else.
 - For those whose interests and research takes them in a statistical direction, this course is a foundation for your own work, but it will typically be supplemented by more advanced courses on particular topics taught either by our department or elsewhere on campus.
- Philosophical overview
 - This course is not a broad "methods" course. It will teach you what to do with numerical data, but it will not really focus on any particular sociological method. Nonetheless, I would like to begin with my own taxonomy of sociological methods, so that you can better understand how statistics fits in.
 - The division between "quantitative" and "qualitative" research is a false one. Its primary flaw is that it muddles the distinction between **data collection** and **data analysis** which are two distinct phases of any sociological method.
 - So, first let's consider the major ways in which sociologists collect data:
 - * **(Participant) observation** - Directly observing a social situation, usually for an extended period of time and often in a participatory manner. Offers a chance for deep insight and rich data. Limited in terms of generalizability and replicability. Also the very richness of data can make data analysis a difficult task.
 - * **Asking questions** - Asking questions of individuals and recording their responses. Comes in two general forms:
 - *Open-ended questioning* - Questions are broad and open-ended and are intended to elicit unique responses from individuals. Open-ended questioning provides richer (and likely more accurate) responses than closed-form questioning. However, the cost of each interview is considerably higher, meaning there are usually far fewer respondents. Like participant observation, the richness of answers can make data analysis difficult.
 - *Closed-form questioning* - Questions are specific and answers generally conform to predetermined categories and ranges determined by the researcher. This form creates the most artificiality in responses, but is easy enough to administer that a large number of people can be questioned and the uniformity of responses simplifies data analysis.
 - * **Archival research** - The collection of data from some historical record. Rarely was the data originally collected for the purposes of the current researcher, so care must be taken in considering the original data collection procedures and how the data itself became available. This form of data collection can often yield novel datasets.

- It is not unusual for a research project to collect data in more than one way. For example, the method of **ethnography** is a wholistic approach to studying a particular social context (originally designed for studying a “people”) which will generally draw on participant observation, questioning (usually open-ended), and sometimes even archival methods.
- Data analysis is the task of processing collected data in order to produce some empirical findings. Typically, it occurs after data collection, although sometimes the two are contemporaneous (most common in participant observation). The methods of data analysis can vary considerably and will depend on two important factors:
 1. The form the data take: field notes, numerically coded responses to a survey, newspaper articles, vital statistics records, interview transcripts, etc.
 2. The question of interest
- Statistics is the analysis of data coded in a numerical format. Although it is most closely tied to closed-form questioning, it can in actuality be applied to any of the data collection techniques listed above.
 - * Record the frequency of some kind of social interaction in a participant observation.
 - * Consider the correspondence between certain words among different respondents in open-ended interviews. (content analysis)
 - * Examine the diffusion of certain laws across states during some historical period. (event history analysis)
- The reason statistics is so closely tied to closed-form questioning is that good surveys are designed in such a way that their **generalizability** to a larger population is quantifiable.
- This leads us to the two major areas of statistics:
 - * **Measurement** - how do we measure things in order to describe data effectively? Most importantly, how do we measure relationships between things?
 - * **Statistical Inference** - How certain can we be that our measurements are generalizable to a larger population?
- Measurement is really the more fundamental of these two, but inference unfortunately often gets a lot more of the glamor.
- In this course we will first focus on measurement of numerical data and then we will move onto the issue of statistical inference.

1.2 The idea of a distribution

- Looking at Data
 - What does the data we use look like?
 - * rows are **observations** (or individuals as IPS call them) Observations can be individual people, states, countries, organizations, etc.
 - * columns are **variables** which take different values across observations.
 - * Variables can be divided into two types:
 1. **Quantitative** variables are simply numbers measuring something (height, dollars, years, etc.). They come in two forms:
 - (a) **Continuous** variables can take on any value within some range of values. Height is a continuous variable.
 - (b) **Discrete** variables take on only certain values, frequently integers. The number of accidents at a factory is a discrete variable.
 2. **Qualitative** (or categorical) variables indicate the category that an observation falls into (race, gender, support for the war in Iraq). They likewise come in two forms:
 - (a) **Nominal** variables are unordered categories. Race is an unordered category.
 - (b) **Ordinal** variables can be ordered in some way. Highest educational degree is an ordinal category: (no degree, HS diploma, AA, BA, post-grad)
 - Our first task will be to summarize information about a single variable. Then we will move on to looking at relationships *between* variables, which is the bread and butter of social science research.
 - We want to understand how the values of each variable are distributed. First we will learn graphical methods to look at the **distribution** of a variable.
 - * Categorical variables
 - Bar Graphs (use example below)
 1. Calculate the **frequency** of each category
 2. Calculate the **proportion** for each category
 3. Plot the proportions or frequencies on a graph (why use proportion?)

Race	Frequency	Proportion
White	267	.722
Black	39	.105
Hispanic	47	.127
Asian and Pacific Islander	17	.046
Total	370	1

· Pie Charts (don't use them)

* Quantitative Variables

· Stem and leaf plots

1. For all the values that variables takes, plot all digits except last in ascending order.
2. Draw a line, and add all single digits for each observation in a row
3. sort these digits into order.
4. Example: Babe Ruth's homeruns (1918-1934)

```

11 29 54 59 35 41 46 25 47 60 54 46 49 46 41 34 22
1 | 1
2 | 259
3 | 45
4 | 1166679
5 | 449
6 | 0

```

5. If there are lots of values then you can "split" into groups of five instead of ten
6. It may be necessary to round in order to do a stem plot, but be careful that rounding does not lose too much of the variation

· Back-to-back stem plots

1. Same as stem-and-leaf plot, but allows comparison of two distributions
2. Simply plot them back-to-back
3. Compare Roger Maris's homerun distribution (1957-1968, excl. 1965)

```

14 28 16 39 61 33 23 26 13 9 5
95 | 0 |
346 | 1 | 1
368 | 2 | 259
39 | 3 | 45

```

```

      | 4 | 1166679
      | 5 | 449
1    | 6 | 0

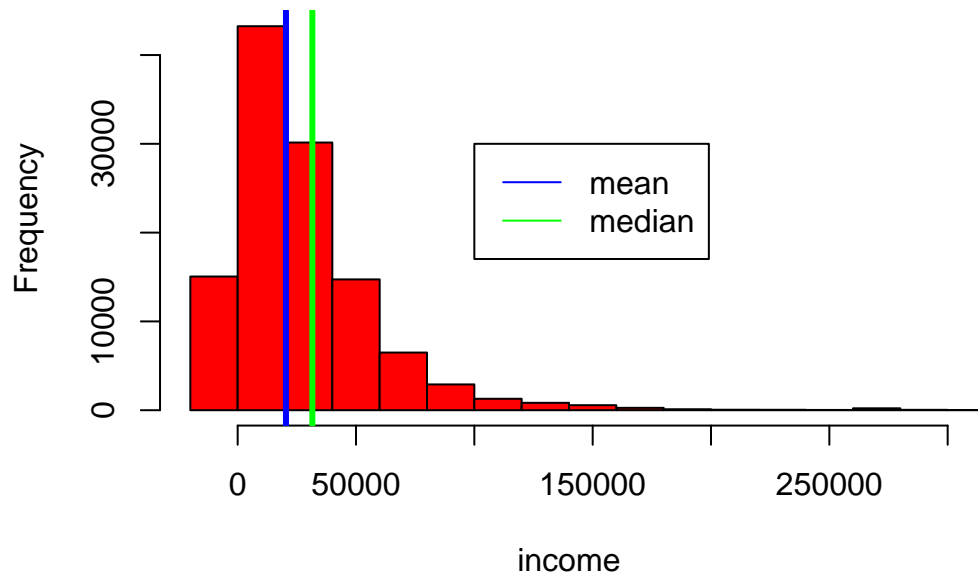
```

4. Distributions must be similar in order to make this possible

· Histograms

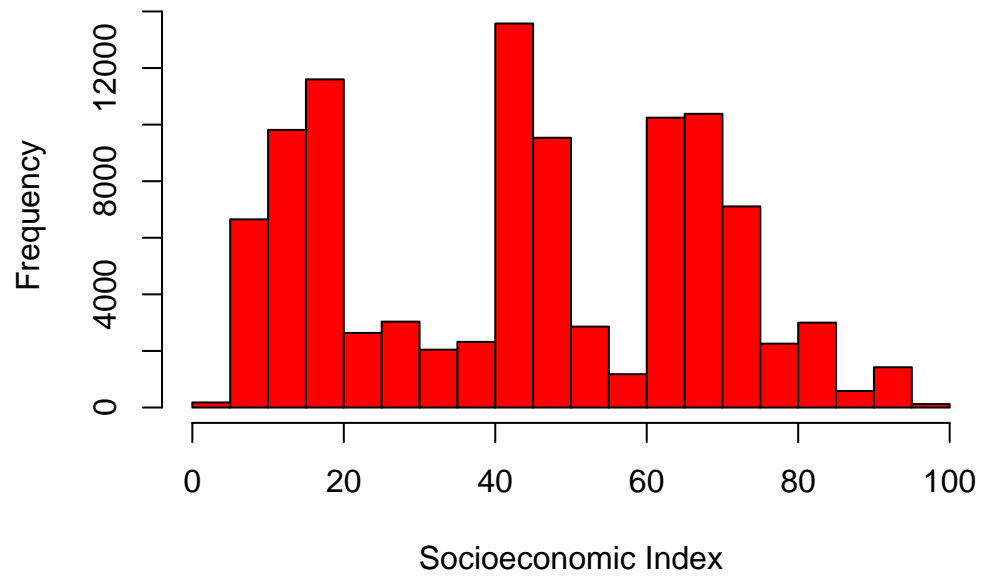
1. When dataset is large, stemplots become cumbersome. We can use a histogram.
2. Break data into *equivalent* intervals (i.e. income 20K-25K, 25K-30K)
3. Count frequency of observations in each interval
4. Plot bars (with no space between) whose width is the interval and whose height is the frequency or relative frequency. Since intervals are the same length, these bars give us a good look at the distribution of the data.
5. Do babe ruth example as a histogram. What intervals to use? Have students count.
6. Go over example with income - what do students notice?

**income distribution of working-age adults
NY Census sample**



7. Go over example with SEI index - what do students notice?

SEI distribution of working-age adults NY Census sample



- What are we looking for in a distribution?
 - * Deviations (**Outliers**) from the pattern
 - * Shape: symmetric or skewed (how about income data?)
 - * Center: how many humps? (unimodal, bimodal, etc.)
 - * Spread: distance between largest and smallest

1.3 Measures of center and spread

- Graphical measures are very good tools for analyzing distributions. But we would like summary measures which can capture the important elements of a distribution.

- We would like a measure for the center of a distribution
- We would like a measure for the spread of a distribution
- Sample of 20 SEI scores for working adults from the 2000 Census: (explain SEI scores)
Convert occupational categories into a composite measure of prestige, based on income and education

27 18 84 44 68
53 51 46 15 44
10 44 22 15 72
73 18 93 68 67

Ordered:

10 15 15 18 18
22 27 44 44 44
46 51 53 67 68
68 72 73 84 93

Stemplot:

```
1 | 05588
2 | 27
3 |
4 | 4446
5 | 13
6 | 788
7 | 23
8 | 4
9 | 3
```

- Review - first let's review some algebra
 - x =female babies at birth (98), y =male babies at birth (102)
 - Ratio - one number over another

$$\frac{y}{x} = \frac{102}{98} = 1.04$$

– Proportion - number out of total

$$\frac{x}{x+y} = \frac{98}{102+98} = 0.49$$

– Percent (per 100), simply multiply by 100

– "Rate" - accounting for exposure in some way. Let's say 2 of the girl babies die in their first year, and 3 of the boy babies die.

$$\frac{2}{98} = 0.02041$$

$$\frac{3}{102} = 0.02941$$

– A true rate always involves a measure of time in the denominator.

– Now let's consider some shorthand:

* summation sign \sum

* subscripting

• Measures of Center (average - don't use this term)

– **mean** - the balancing point of a distribution (draw picture)

$$\bar{x} = (x_1 + x_2 + x_3 + \dots + x_n)/n$$

or

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

Example: $\sum x_i = 932$, $\bar{x} = 932/20 = 46.6$

– **median** - the midpoint of a distribution, so that half of the observations are to the left and half to the right. (draw a picture)

1. Re-order observations in order.
2. If n is odd, then find the midpoint value in this ordered list (i.e. 3 of 5)
3. If n is even, then find the two midpoint values and average them.

Example: the 10th observation is 44, the 11th observation is 46

$$\frac{44+46}{2} = 45$$

– **mode** - the most common observation, i.e. "the peak". We have already talked about this in terms of whether the distribution is "unimodal." We will not use the mode very often, but you should know it. (Example=44)

- Have students go through some sample distributions and try to locate these things.
 1. symmetric, normal distribution
 2. left-skewed
 3. right-skewed
 4. bimodal, symmetric
- What do you notice, mean is pulled more by extreme values than median. In general, mean is more sensitive than median. (show income distribution)
- Measures of Spread - choice of measure depends on choice of measure of center
 - Using Median
 - * We could measure the **range** - the distance between the smallest and largest point, but this would not be very useful because of outliers.
 - * Better to make use of the **quantiles** (or percentiles) of the distribution. The median is basically the 50th percentile of the distribution.
 1. sort the observations so they are in ascending order.
 2. For the p th percentile, find the ordered observation which corresponds to that observation.

$$\frac{i}{n} * 100 = p$$
 3. For empirical distributions, you will often have to find the closest value rather than the exact value.
 4. This value is the p th percentile of the distribution
 5. We are often interested in the **quartiles** of the distribution: the 25th, 50th, and 75th percentiles.
Example: $\frac{18+22}{2} = 20$, $\frac{68+68}{2} = 68$
 6. Taking the difference between the 75th and 25th percentile gives us the **interquartile range (IQR)**, which is a measure of the spread of the distribution, which is less affected by outliers than the range.
Example: $68 - 20 = 40$
 - * Using the IQR, we can now get a five-number summary of the distribution:

Minimum	25th	median	75th	Maximum
10	20	45	68	93
 - * We can use this five number summary to produce another important kind of graph called the **boxplot**. (go through steps with the example)

1. On the y-axis draw a box which extends from the 25th to the 75th percentiles.
2. put a line through the box to indicate the median.
3. put whiskers extending from the box to the minimum and maximum, unless these values are higher than some rule of thumb from the median, (1.5xIQR).
4. plot points higher than rule of thumb individually.

– Using Mean

- * When spread is measured relative to the mean, we use something called the **variance**(s^2) and its square root the **standard deviation** (s).

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

- * Go through this step at a time
 1. How large the values of $x_i - \bar{x}$ are on average gives us some measure of how spread out the observations are around the mean. But we can't average these values because they will sum to zero (show this)
 2. So first we will square these values, so that they are all positive.
 3. Then we will sum them up.
 4. We want some measure of "average" squared distance from the mean, so we divide by the number of observations. However, we have to subtract one from this number first, because we used the mean to calculate the variance, we have one less **degree of freedom**
 5. This measures variance, but our units are now squared because of the squared term earlier, so we take the square root of this to get a measure which is in the same units as the mean.
 6. Go through example

x_i	$x_i - \bar{x}$	$(x_i - \bar{x})^2$
27	-19.6	384.16
53	6.4	40.96
10	-36.6	1339.56
73	26.4	696.96
18	-28.6	817.96
51	4.4	19.36
44	-2.6	6.76
18	-28.6	817.96
84	37.4	1398.76
46	-0.6	0.36
22	-24.6	605.16
93	46.4	2152.96
44	-2.6	6.76
15	-31.6	998.56
15	-31.6	998.56
68	21.4	457.96
68	21.4	457.96
44	-2.6	6.76
72	25.4	645.16
67	20.4	416.16
		$\Sigma = 12268.8$

$$s^2 = \frac{12268.8}{20 - 1} = 645.7263$$

$$s = \sqrt{645.7263} = 25.41115$$

* Like the mean, s is sensitive to outliers and skewness.

1.4 Measuring relationships with categorical variables

- Introduction to Relationships
 - What we often care about in sociology is whether there is a relationship between "things"
 1. Does the educational level of parents affect their children's educational attainment?
 2. Does income inequality affect mortality rates?
 3. Does divorce affect children's mental well-being?
 - There is a difference between association and causation:
 - * association: two things tend to go together
 - * causation: one thing causes the other
 - In the latter case, we often refer to one variable as **independent** (explanatory) and the other as **dependent** (response). We are interested in how variation in the dependent variable might be explained by differences in the explanatory variable.
 - But first we begin with association
- We are going to use a new dataset to examine the relationship between categorical variables. We are going to use data on the passengers on the Titanic. We have several pieces of interesting information:
 - Whether the passenger survived
 - The age of the passenger
 - The class of the passenger (1st, 2nd, 3rd)
 - The sex of the passenger
 - Number of relatives traveling with person
- Measuring Association, 2 categorical relationships
 - When both variables are categorical, the best approach is to calculate the conditional distribution of one of the variables. This requires a fair amount of calculation. You should start with a two-way table.

- Let's look at the two-way table of class by survivorship.

	Survived	Died
1st Class	200	123
2nd Class	119	158
3rd Class	181	528

- The cells in this table give the **joint distribution** of these two variables. We will define a term f_{ij} to refer to the count of observations in the i th row and the j th column.
- We want to know whether class is related to survivorship. The most straightforward way to go about this is to calculate the distribution of one variable conditional on being in a certain category of the other one. This is called a **conditional distribution**. It doesn't technically matter which variable we condition on, but for our purposes it makes intuitive sense to look at the distribution of survivorship conditional on class.
- In order to calculate the conditional distribution, we first need to calculate the **marginal distributions**. These are the distributions of each variable regardless of the other and can be obtained by summing across the rows and columns.

	Survived	Died	Total
1st Class	200	123	323
2nd Class	119	158	277
3rd Class	181	528	709
Total	500	809	1309

We represent marginal distributions with dots to indicate what was summed over:

$$f_{i.} = \sum_{j=1}^J f_{ij}$$

$$f_{.j} = \sum_{i=1}^I f_{ij}$$

- In order to calculate conditional distributions, we divide through by either the column or row totals. In our case, we want the conditional distribution of survivorship, so we are going to divide through by the row marginals.

	Survived	Died	Total
1st Class	200/323=0.62	123/323=0.38	1
2nd Class	119/277=0.43	158/277=0.57	1
3rd Class	181/709=0.26	528/709=0.74	1

In mathematical terms, the distribution of survivorship given being in first class is:

$$\frac{f_{1j}}{f_{1.}}$$

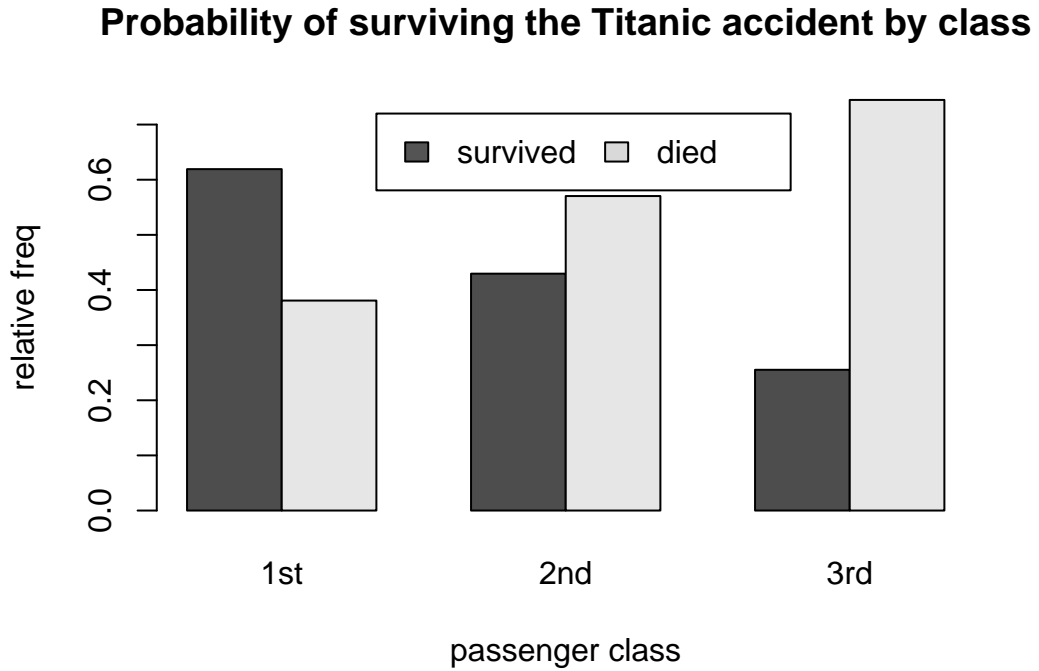
given being in second class:

$$\frac{f_{2j}}{f_{2.}}$$

given being in third class:

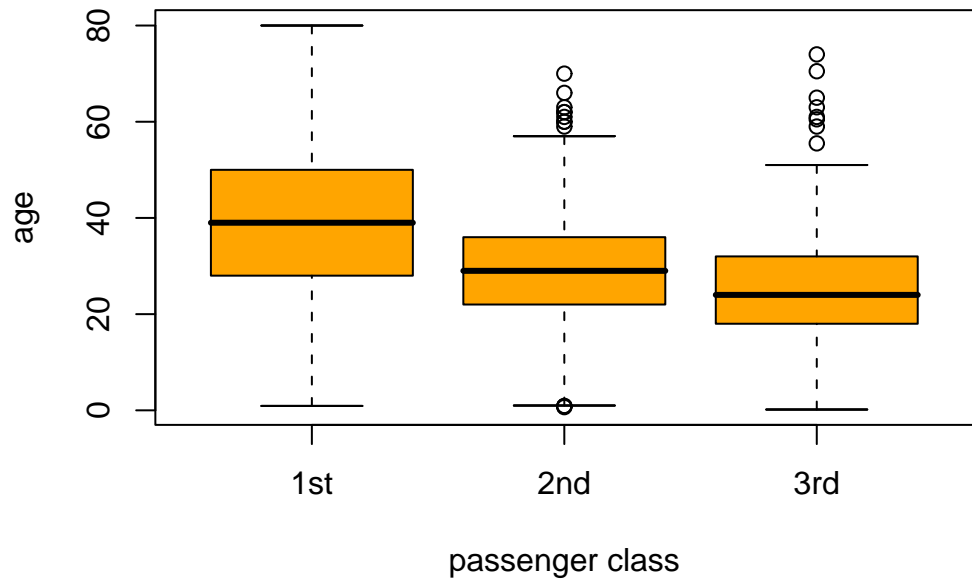
$$\frac{f_{3j}}{f_{3\cdot}}$$

- In order to see this relationship graphically, let's graph the three conditional distributions.

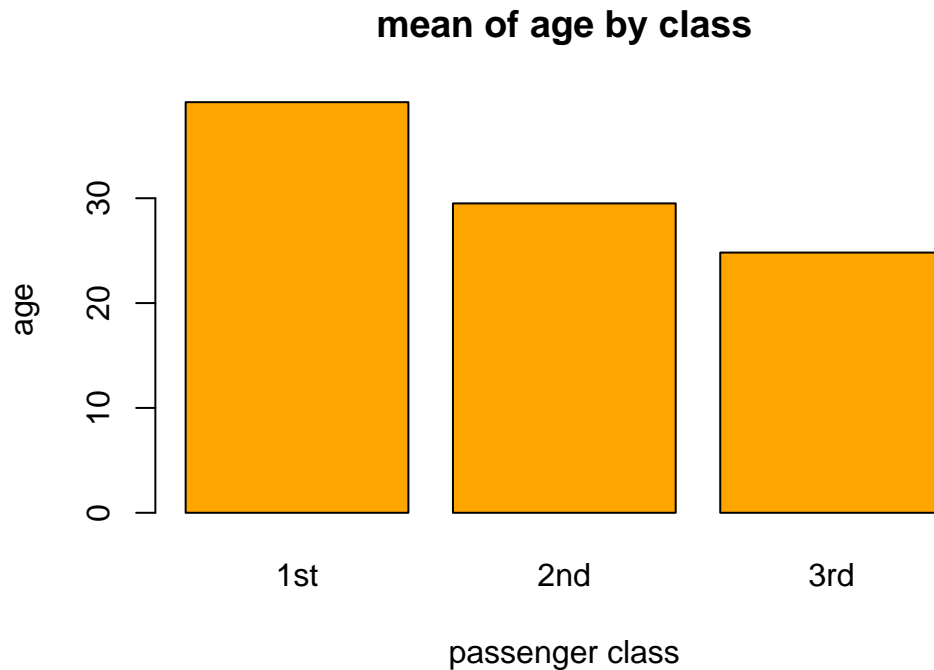


- These conditional distributions are clearly different. The higher the class of the passenger, the more likely they were to survive the Titanic.
- Measuring association (1 quantitative and 1 categorical variable)
 - Let's examine the relationship between age and passenger class in the Titanic data.
 - When you want to examine the relationship between a quantitative and categorical variable, use:
 - * a side-by-side stemplot (cumbersome if more than two categories), or
 - * multiple boxplots

boxplots of age by class



* Bargraphs



- If you only have two categories, you could measure the mean difference between them. In the titanic data the mean age of survivors (\bar{x}_s) was 30.55 and the mean age of those who died was (\bar{x}_d) 28.92, so:

$$\bar{x}_s - \bar{x}_d = 30.55 - 28.92 = 1.63$$

So, survivors were slightly older than non-survivors. Any idea why this might be so?

- Describing association: if category is not ordered than you can't speak of association as positive or negative, just different or not different.
- Be careful about interpreting results. It is possible for two variables to be related but not necessarily causal. For example, we may think that older people were more likely to survive because they were first class passengers, rather than because being older itself was more advantageous. This is the problem of **lurking variables** which constantly bedevils us (draw a picture).
- A classic example of this problem comes from data on admissions to six graduate programs at the University of California, Berkeley in 1973.

- The admissions data clearly show that men were more likely to be admitted than women.

	Male	Female
Admitted	1198	557
Rejected	1493	1278

calculate the distribution of acceptance by gender:

	Male	Female
Admitted	0.45	0.30
Rejected	0.56	0.70

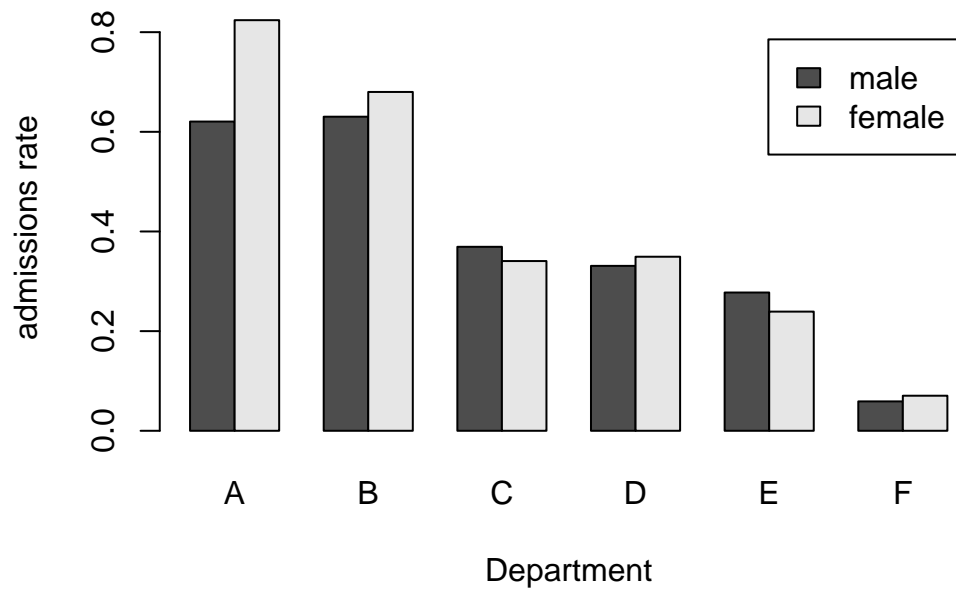
- The problem is that each program has its own admissions practices, so let's look at the gender-conditional acceptance rate for each program separately.

	Dept A		Dept B		Dept C		Dept D		Dept E		Dept F	
	M	F	M	F	M	F	M	F	M	F	M	F
Admitted	512	89	353	17	120	202	138	131	53	94	22	24
Rejected	313	19	207	8	205	391	279	244	138	299	351	317

Let's calculate the gender conditional acceptance for each program.

	Dept A		Dept B		Dept C		Dept D		Dept E		Dept F	
	M	F	M	F	M	F	M	F	M	F	M	F
Admitted	0.62	0.82	0.63	0.68	0.37	0.34	0.33	0.35	0.27	0.24	0.06	0.07

And let's plot out these acceptance rates by gender for each department.

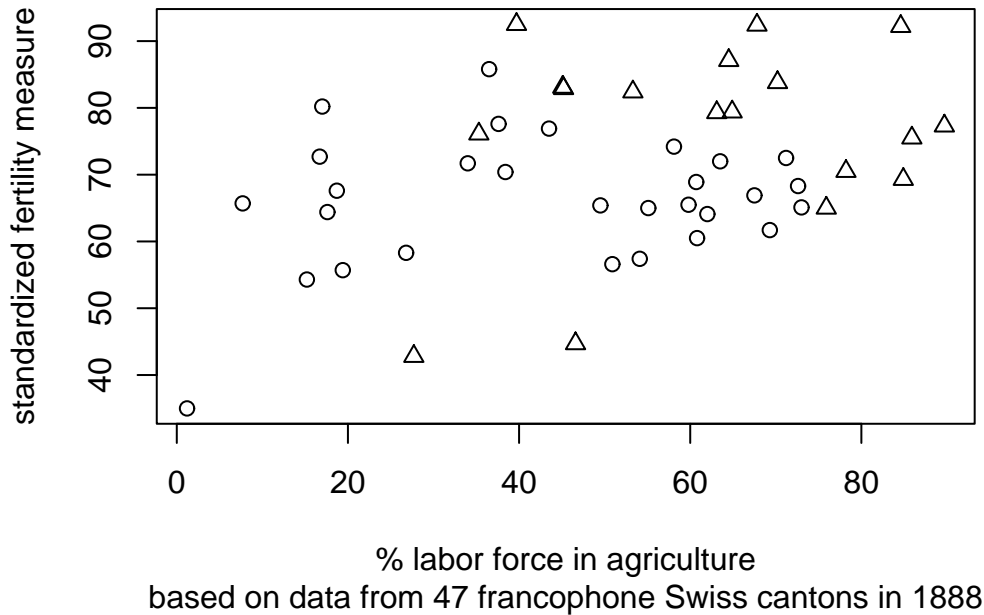


- The results clearly show that men do not have an advantage over women in the admissions practices of any department. In fact, the only gender disparity goes in the opposite direction, favoring women in Department A.
- Why the discrepancy? Women are more likely to apply to departments with higher rejection rates.

1.5 Measuring relationships between quantitative variables

- Graphing Association, 2 quantitative variables
 - When both variables are quantitative, we can use the scatterplot to graphically depict the association between the two variables.
 1. Each point is an observation
 2. independent variable on the x-axis
 3. dependent variable on the y-axis
 - As an example, let's look at the relationship between fertility and agriculture in Switzerland. (Discuss demographic transition and the Princeton Fertility Project)

Scatterplot of agriculture and fertility in Switzerland (1881)



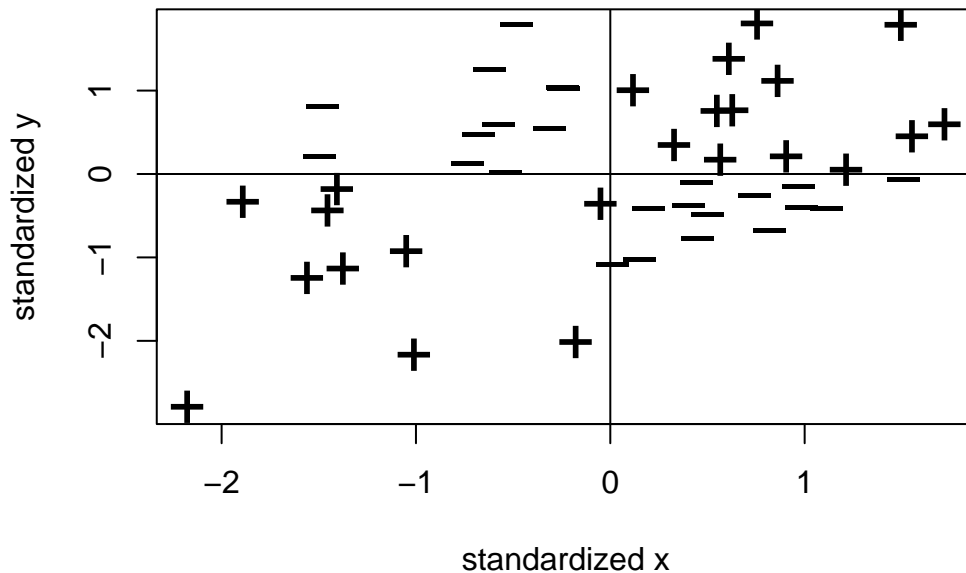
- What to look for in a scatterplot
 1. Is there a relationship? How strong is this relationship?
 2. What is the direction of this relationship? **positive** or **negative association**
 3. Is this relationship linear - do the points follow a roughly straight line or is there a more complex pattern?

4. Are there any obvious outliers to the general pattern?

- Graphical analysis can give you an idea about relationships, but scatterplots can be misleading about the strength of association by changes in scale
- We want a numerical measure of the strength and direction of the association. We use a measure called the **correlation coefficient**, or r :

$$r = \frac{1}{n-1} \sum \frac{x_i - \bar{x}}{s_x} * \frac{y_i - \bar{y}}{s_y}$$

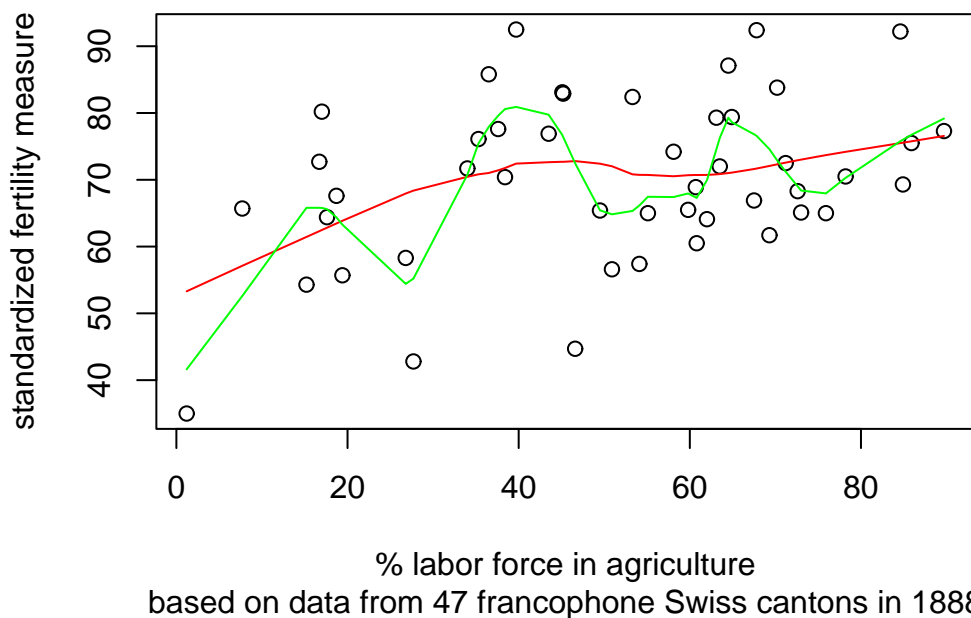
- Let's go through this formula, one element at a time (show with example, graphically, plot out mean lines of x and y)



1. We have seen $\frac{x_i - \bar{x}}{s_x}$ and $\frac{y_i - \bar{y}}{s_y}$ before. The values of x and y have just been standardized so that they are measured on the same scale.
2. Next, we take the product of these standardized values. If an individual i has values on both these variables that are either above or below the mean, then their product will be positive. If this individual has a positive value for one and a negative value for the other, however, then their product will be negative.

3. Next, we sum the products. Positive values will increase this sum, while negative values will decrease it. If lots of individuals had a positive association among these variables, then this number will get large. If lots of individuals had negative associations, then this number will be highly negative. If individuals with positive and negative associations are balanced, then the number will be close to zero.
 4. Finally, we take the "average" of this sum, to factor out sample size.
- Properties of r
 1. which variable is x and which is y is irrelevant - only measures association
 2. r is unit-less
 3. a positive r indicates positive association
 4. a negative r indicates negative association
 5. r is always between 1 and -1. 1 and -1 indicate perfect correlation (straight lines), while 0 indicates no association.
 6. r can be affected by outliers.
 - In our example, r is 0.35. This is a moderately strong and positive relationship.
- Putting lines through points: The crude truth about most social science statistical research is that we are conceptually putting lines through a set of points, in order to predict one variable from another. There are multiple ways to do this:
 - **Smoothing.** Not all relationships are linear. Rather than forcing a straight line through a set of points, we could take a milder approach and try to "smooth" the points to see if a trend shows up.
 - * There are different methods for smoothing. The basic idea of each one is the same.
 - * Rather than taking the values of y for each point you use all the points within some range of x to calculate a smoothed value of y (draw a picture). The methods only differ in how they do the estimation.
 - * Running means compute the mean of the surrounding y variables.
 - * Running medians compute the median of the surrounding y variables.
 - * LOWESS uses a more complicated technique involving weighted regression to predict the value of y .
 - * Example: use LOWESS smoothing on the scatterplot.

Scatterplot of agriculture and fertility in Switzerland (1888) with lowess smoothers (f=.66 and f=.2)



- **OLS Regression.** Another more common way to do this would be to fit a straight line through our cloud of points.

$$y = b_0 + b_1x$$

For our purposes, we want to predict the value of y_i from x_i , so our formula is:

$$\hat{y}_i = b_0 + b_1x_i$$

* Review meaning of linear equation

1. b_0 is the **intercept** - the point on the y-axis the line crosses when $x=0$
2. b_1 is the **slope** - the increase (or decrease) in y for every unit increase in x .

* What is the best way to draw a line through these points?

- We want to minimize the overall vertical distance of points from the line (draw picture, show vertical distances) - vertical distance because we are trying to minimize variation in y . The sum of this overall squared distance (SSR) which we want to minimize is given by:

$$\sum (\hat{y}_i - y_i)^2$$

- We use the **ordinary least squares regression** (OLS) technique, which minimizes the sum of the squares of these distances.
- We are not going to derive how this technique works, but it turns out that minimizing these squared sums gives you:

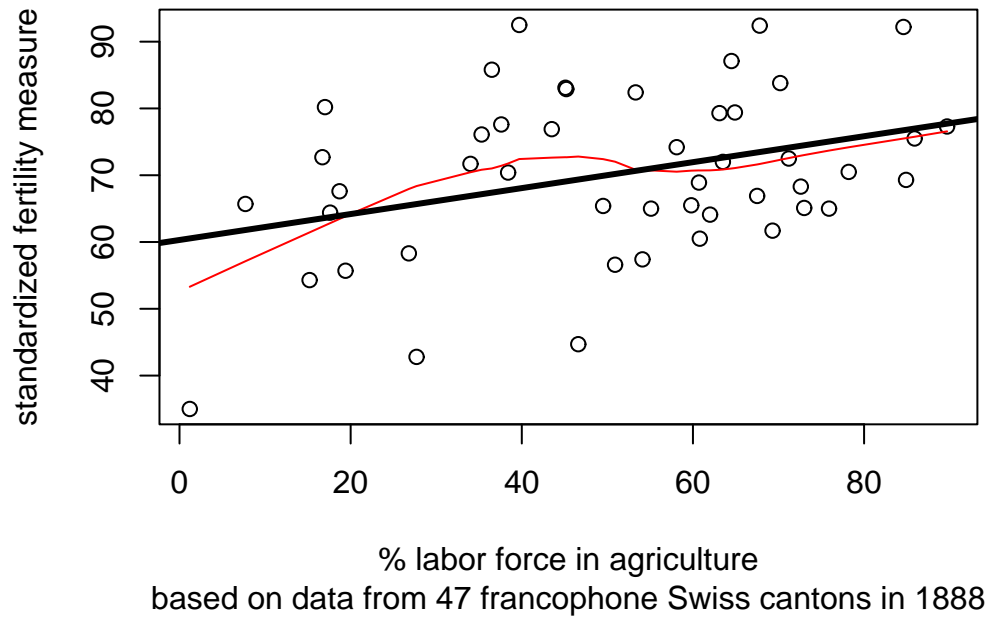
$$b_1 = r * \frac{s_y}{s_x}$$

The most important point here is that b_1 is just a scaled version of r .
 The OLS line always passes through (\bar{x}, \bar{y}) , so:

$$b_0 = \bar{y} - b_1 * \bar{x}$$

- In our example $b_0 = 60.3$ and $b_1 = 0.19$. Show graphically:

Scatterplot of agriculture and fertility in Switzerland (1881) with OLS regression line and lowess smoother (f=.66)



- Once again, we must be careful how we interpret results because there could be lurking variables. Discuss issue of education in relation to the swiss fertility results.

1.6 Interpreting OLS regression and transformations

- As an example today, we will be using opinion responses to an abortion question on the General Social Survey (1994) as a dependent variable. The specific opinion statement is "A pregnant woman should be able to obtain a legal abortion for any reason whatsoever, if she chooses not to have the baby." Respondents could strongly agree (5), agree (4), neither agree or disagree (3), disagree (2), and strongly disagree (1).
- Let's attempt to predict abortion attitudes by years of education.

We have the following information:

$$r = 0.155, \bar{y} = 2.96, s_y = 1.48, \bar{x} = 13.23, s_x = 2.89$$

So,

$$b_1 = r \frac{s_y}{s_x} = (0.155) \frac{1.48}{2.89} = 0.079$$

$$b_0 = \bar{y} - b_1 \bar{x} = 2.96 - 0.079 * 13.23 = 1.91$$

Variable	Coefficient
Intercept	1.91
Education (years)	0.079

What are the meanings of b_0 and b_1 ?

- The model predicts that for every year of education, support for abortion will increase by a score of 0.079.
- The model predicts that those with no education will have an average attitudinal score of 1.91.

Plot this on a graph using zero years of education and 20 years of education:

$$0 \text{ years of education: } \hat{y} = 1.91 + 0.079 * 0 = 1.91$$

$$20 \text{ years of education: } \hat{y} = 1.91 + 0.079 * 20 = 3.49$$

- How do we assess the quality of x as a predictor of y ?
 - We can ask the question: How much of the overall variance of y is explained by x ?
 - Draw a picture. Plot mean line, and then OLS line and show that the residuals should be smaller.
 - Measure the overall variance in residuals from the mean line:

$$SSY = \sum (y_i - \bar{y})^2$$

- Measure of the overall variance in residuals from the OLS regression line:

$$SSR = \sum (y_i - \hat{y})$$

- What proportion of the overall variance of SSY is in SSR?

$$\frac{SSR}{SSY} = \frac{\sum (y_i - \hat{y})}{\sum (y_i - \bar{y})}$$

- By what proportion have we reduced SSY by using x as a predictor?

$$1 - \frac{SSR}{SSY}$$

- It turns out that this proportion is given by the square of r : r^2 .

- * $r = 1, r^2 = 1$

- * $r = 0.707, r^2 = 0.5$

- * $r = 0.574, r^2 = 0.33$

- * $r = 0.5, r^2 = 0.25$

- * $r = 0.3, r^2 = 0.09$

- * $r = 0, r^2 = 0$

- In our example, the r value is 0.155, so r^2 is 0.024. Years of education only explains 2.4% of the variance in abortion attitudes. Although this may seem small, in large surveys of individuals, this is often pretty typical.

- Mean-centering variables

- The intercept always represents the predicted value of the dependent variable when all of the independent variables are zero. This can often be a meaningless value because the realistic range of some independent variables does not include zero (think about including a year variable in a regression analysis to get a time trend - the zero value would be 0 AD).

- This can be particularly problematic when adding variables sequentially to models because the intercept can change significantly simply because of the range of added variables.

- The simple solution to this problem is to shift the distribution of independent variables to make the zero point more interpretable. There are two common solutions:

1. Subtract the mean from each independent variable (mean-centering). In this case the intercept will give you the overall mean for the dependent variable.

2. Subtract some substantively meaningful value. For example, if you had a year variable, you might like the intercept to reflect the mean for a particular year.

– Example with abortion data.

Variable	Original	w/mean centering
Intercept	1.91	2.96
Education (years)	0.079	0.079

• **Transforming variables.** What do you do if your relationship is non-linear? Luckily, all hope is not lost. Many non-linear relationships can be turned into linear relationships by transforming one or both variables.

– If you have some idea of the correct functional form between variables, then transforming a non-linear relationship is often straightforward.

* An exponential relationship

$$y = ae^{bx}$$

take the (natural) log of both sides:

$$\log(y) = \log(a) + bx$$

So running a regression of x and $\log(y)$ will fit an exponential curve on the original scale. b is the growth rate and the starting value a is given by the exponential of the intercept.

* Power law (body size to brain size)

$$y = ax^p$$

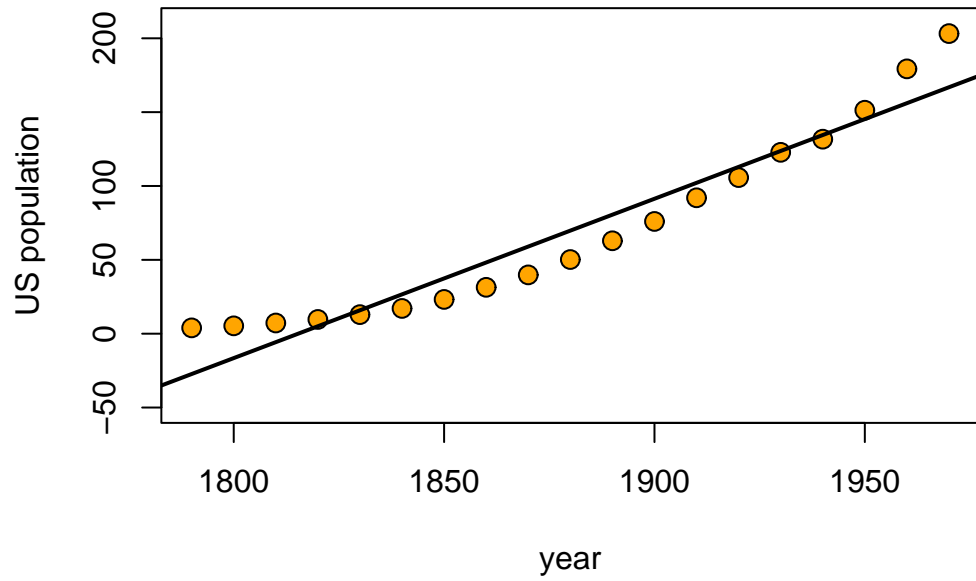
Take the log of both sides again:

$$\log(y) = \log(a) + p \log(x)$$

So running a regression of $\log(x)$ and $\log(y)$ will fit a power law curve on the original scale. p is given by the slope.

* As an example, let's look at the size of the US population from 1790 to 1970.

US population growth



Population growth is frequently fit better by exponential growth curves:

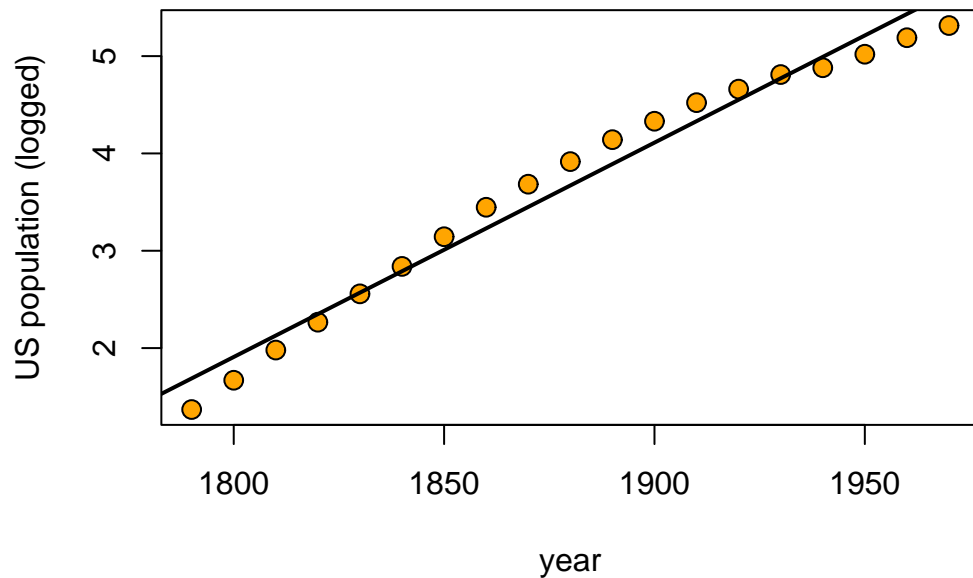
$$P(t) = P_0 e^{bt}$$

If we log this equation, we get:

$$\log(P_t) = \log(P_0) + bt$$

So a transformation is in order. Let's log our population and look at the relationship now:

US population growth (log-scale)

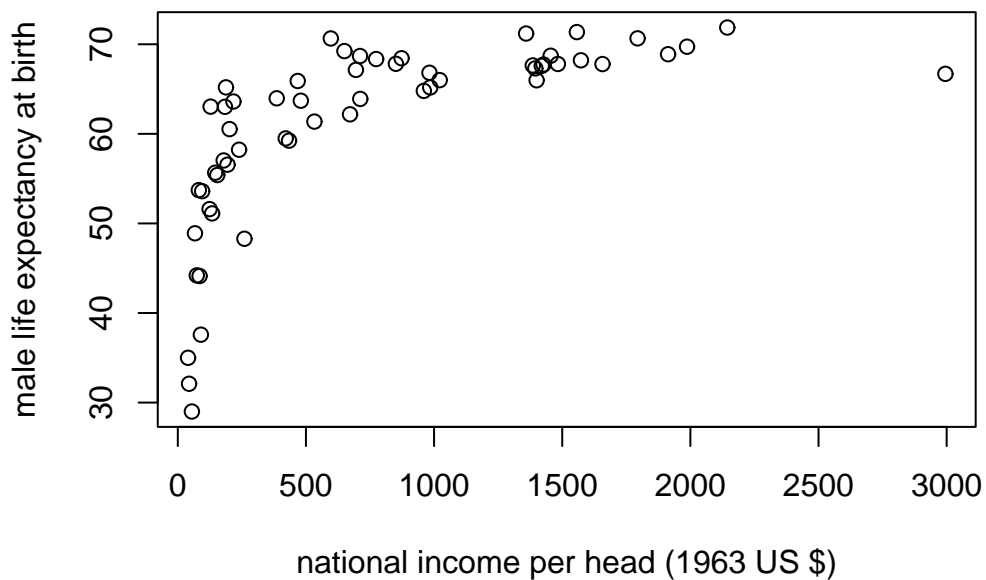


The relationship looks much more linear. Using this model, we get an exponential growth rate of 0.022. Note that there is still some slight curvature to the line. Any idea what is going on here?

- Good transformations are **monotonic**. That is, they preserve the ordering in the data (biggest remains biggest) or invert the ordering exactly (biggest becomes smallest).
- It is sometimes useful to transform independent variables in order to reduce skewness and the effect of outliers. When $p < 1$, transformations will pull in right tails and push out left tails. When $p > 1$, the opposite is true. The most common method is to log the variable to reduce right skewness (on income, for example).
- Keep in mind that such transformations also change the shape of the relationship. Taking the natural log of the independent variable, for example, implies a "diminishing returns" relationship.

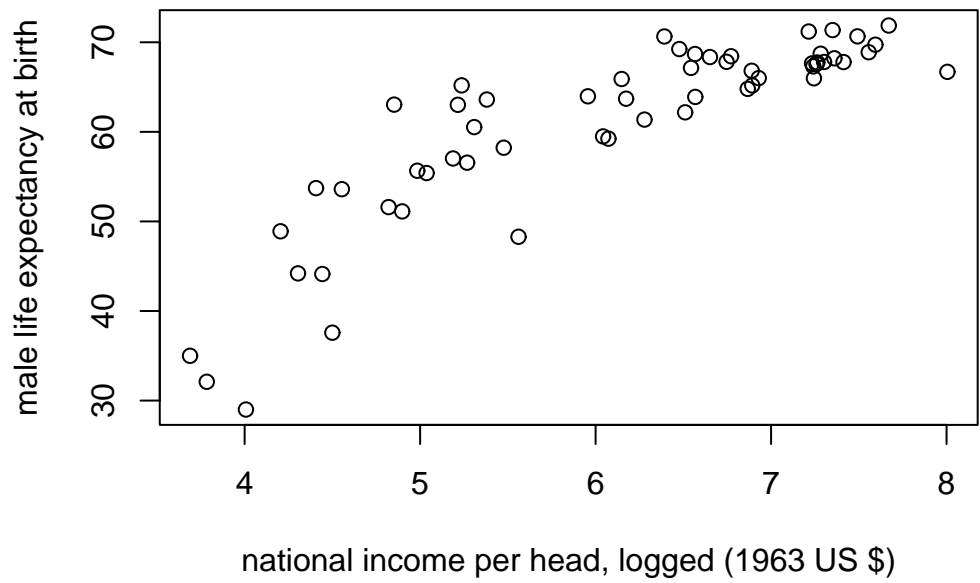
Let's look at an example of this kind of relationship. The graph below shows the scatterplot of national income per head and life expectancy for 57 countries in 1960.

scatterplot of life expectancy and national income 1960



The pattern here is clearly non-linear and suggests a diminishing returns model. Increases in national income matter more for life expectancy when national income is low. We can better model this relationship by logging our national income variable.

scatterplot of life expectancy and national income on the log scale, 1960



- Another common method is to use a polynomial expression of x on y in order to pick up curvature in the relationship. We will discuss this later in the term, when we learn how to add more than one predictor of y in the same equation.

1.7 Multivariate regression

- Multiple Regression

- Before we had: (review)

$$\hat{y}_i = b_0 + b_1x_i$$

- Now, let's add another variable:

$$\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2}$$

- Think of a three-dimensional space (draw this space)
- We are still finding the line that minimizes:

$$\sum (y_i - \hat{y}_i)^2$$

- What if we add even more variables?
- Imagine a p-dimensional space:

$$\hat{y}_i = b_0 + b_1x_{i1} + b_2x_{i2} + b_3x_{i3} + \dots + b_px_{ip}$$

- We are fitting a line through this p-dimensional space. We are predicting y by all of these x variables together.
- r^2 becomes R^2 . Meaning is the same.
- What is the advantage of doing this?
 - * To "control" for the effects of other variables
 - * The "effect" b_j of one variable is the effect after controlling for the other variables
 - * A definition of b_1

Holding the variables of x_2 through x_p at the same level, the model predicts that a one unit change in x_1 will on average lead to a b_1 unit change in y .

The second part of this definition is the same. It is the first part which has been added.

- * When two explanatory variables are correlated with one another, then including one but not the other in a regression will pick up an indirect effect of the other variable on y . (show with a picture)
- * Example: predict abortion attitudes by age and education separately and then together

Coef.	Model 1	Model 2	Model 3
Constant	2.96	2.96	2.96
Age	-0.010		-0.007
Education		0.079	0.071
R^2	0.012	0.024	0.031

- * What is going on? (Go through it with students)
- * If two independent variables are **orthogonal**, then their b 's are the same whether they are included in a regression together or separately. This occurs when r is zero. Experiments are designed to be orthogonal, but very few variables in real-world situations are orthogonal.
- * Multiple regression is useful for testing or pre-empting claims of **lurking variables**. (draw a picture)

- Multiple regression tricks of the trade

- Including two-category dummy variables

- * y is attitudes towards abortion (1 to 5)
- * x_1 is a **dummy** variable (or indicator variable) indicating whether the individual was male or female (male as reference) (show coding)
- * x_2 is years of education.
- * Put these into a regression equation

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \epsilon_i$$

Variable	Coefficient
Intercept	2.93
Education (years)	0.080
Female	0.060
R^2	0.024

- * β_2 should be easy to interpret as the change in attitude associated with one more year of education. But what is the interpretation of β_1 ?
- * Let's separate this into two equations. (draw them on the board)

Men:

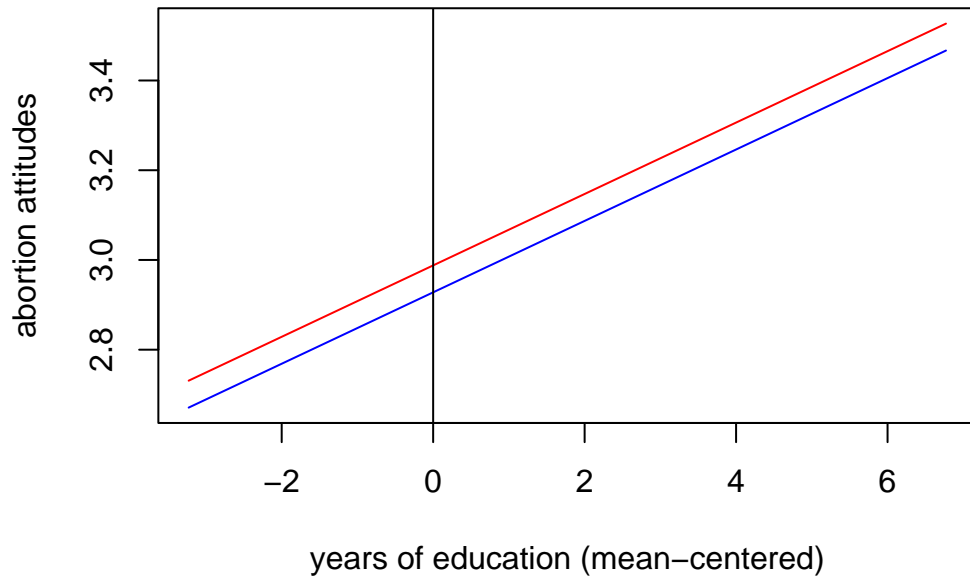
$$y_i = \beta_0 + \beta_2 x_{i2} + \epsilon_i$$

Women:

$$y_i = (\beta_0 + \beta_1) + \beta_2 x_{i2} + \epsilon_i$$

- * The difference is a different intercept. β_1 is the difference in the mean level of support for abortion between men and women.
- * The implication is that gender has the same effect on support for abortion *at all education levels*. The lines are parallel. This model assumes that the relationship between education and support for abortion does not vary by gender. Show this with a graph:

predicted abortion attitudes by education and gender no interaction



- Including categorical variables with more than two categories.
 - * Example: We might think marital status will affect attitudes toward abortion. (married, widowed, divorced, never married)
 - * We might be interested in how these groups differ with respect to some outcome variable. We can include them in a regression model by:
 1. Code each category as a dummy variable (go through this)
 2. put all of these dummy variables **except one** into the regression equation.

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \epsilon_i$$

- * What are these equations for the four groups:

- Married: $y_i = \beta_0$
- Widowed: $y_i = \beta_0 + \beta_1$
- Divorced: $y_i = \beta_0 + \beta_2$
- Never Married: $y_i = \beta_0 + \beta_3$

Variable	Coefficient
Intercept	2.88
Widowed	-0.238
Divorced	0.123
Never Married	0.442
R^2	0.018

- * The intercept is the mean of y for whites. β_1 is the difference between the mean of whites and blacks. β_2 is the difference in means between American Indians and whites.
- * The omitted category is the **reference group**. All the other β 's refer to it. Which category you make the reference group is arbitrary and will depend on what you are interested in looking at.

– Interaction terms

- * What if we think the relationship between x and y is affected by which group the person is a member of? Then we can use interaction terms.
- * Go back to abortion example. Suppose:

$$x_3 = x_1x_2$$

And we have a new equation:

$$y_i = \beta_0 + \beta_1x_{i1} + \beta_2x_{i2} + \beta_3x_{i3} + \epsilon_i$$

Variable	Coefficient	w/o education centering
Intercept	2.93	2.15
Education (years)	0.059	0.059
Female	0.060	-0.490
Education*Female	0.042	0.042
R^2	0.026	

- * Let's separate this into two equations as before. (draw them on the board)
- Men:

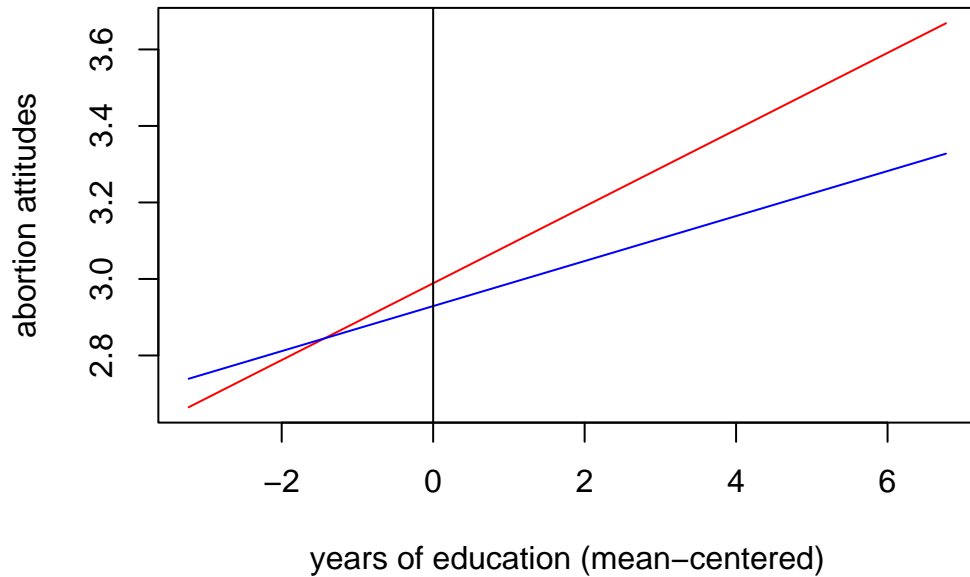
$$y_i = \beta_0 + \beta_2x_{i2} + \epsilon_i$$

Women:

$$y_i = (\beta_0 + \beta_1) + (\beta_2 + \beta_3)x_{i2} + \epsilon_i$$

- * Before women just had a different intercept from men. Now they have a different intercept and slope. β_3 tells us how different the slope is for women relative to men.

predicted abortion attitudes by education and gender interaction term



- * Note that if education is not mean-centered the effect of female in an interaction term is hard to interpret.

– Polynomial Regression

- * What is the 1st assumption of OLS regression? Linearity.
- * Non-linear relationships can often be described by polynomial equations.

$$y = a + bx + cx^2$$

Does everyone remember this equation? We could make it even more complex with a cube:

$$y = a + bx + cx^2 + dx^3$$

- * The more terms you add, the more inflection points you will get in your curve.

- * If we think that the relationship between y and x in our data follows this form, we can model it using multiple regression simply by taking x to that power and including it.

$$x_{i2} = (x_{i1})^2$$

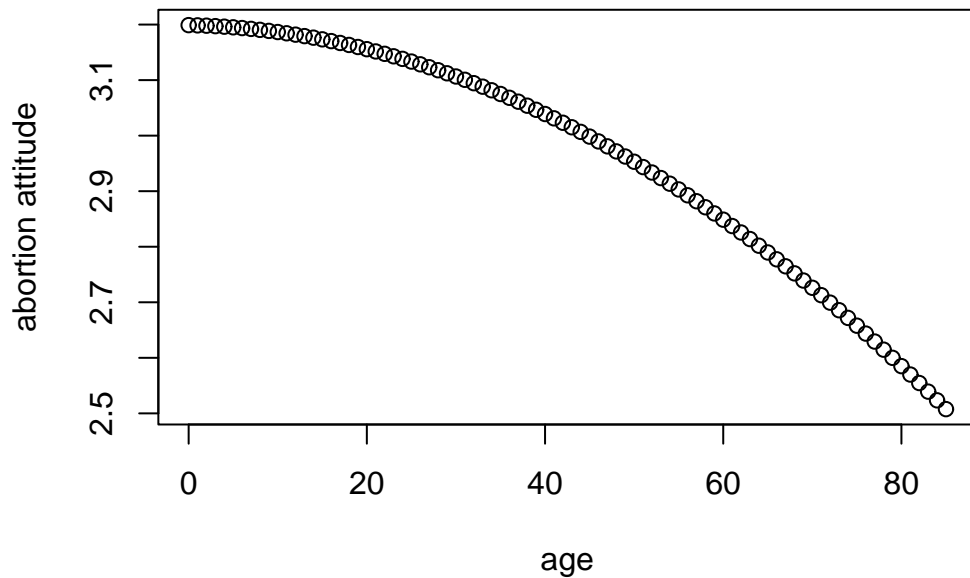
$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2}$$

- * The β 's can be difficult to interpret directly. Best approach is to graph it.
- * Let's think about the effect of age in abortion attitudes. The effect across birth cohorts may be non-linear.

Variable	Coefficient
Intercept	2.99
Age (mean centered)	-0.00877
Age squared	-0.00009
R^2	0.013

Plot this effect:

predicted abortion attitudes by age



The strength of association seems to be diminishing somewhat with age.

– Let's now put together the full example

Variable	Coefficient
Intercept	2.82 (39.8)
Age (mean centered)	-0.0033
Education (mean centered)	0.056
Female*Education	0.035
Sex	
Male (ref)	-
Female	0.089
Marital Status	
Married (ref)	-
Widowed	-0.050
Divorced	0.134
Never Married	0.394
R^2	0.042

1.8 A real research example

- Oppositional Culture Paper
 - What is the theory that the Authors are taking issue with?
 - * Ogbu's Oppositional Culture Theory - i.e. Acting White.
 - Applies to "non-voluntary" migrants (African Americans and American Indians)
 - These groups perceive they have limited job opportunities
 - Because of this perception, they exert less effort in school
 - They come to see school success as "acting white" or "uppity" - thus minority children with high aptitude face pressures to underperform
 - What are four main tenets according to Arnsworth-Darnell and Downey?
 - What data do the authors use to test these tenets?
 - How do they categorize their groups?
 - Go through each hypothesis and hit points noted in my copy of the paper.
 - Discuss conclusions

1.9 Regression diagnostics and cautions: outliers and influential points

- Regression is a powerful and common tool in the social sciences, but it must be used carefully and thoughtfully, or it can produce misleading results.
- There are many diagnostic checks one can make to assess the validity of regression assumptions.
- Cautions

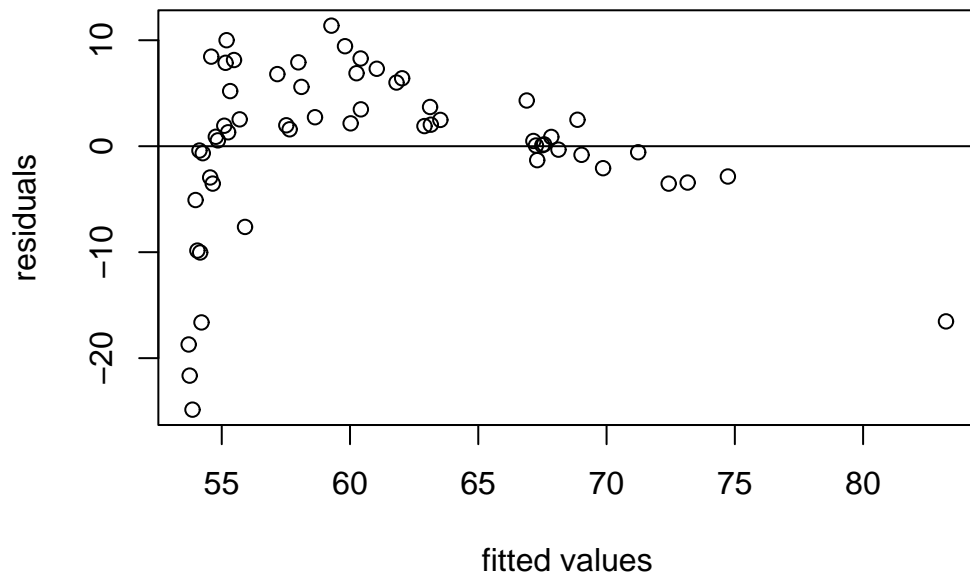
1. Regression only applies when the relationship between variables is linear.

- Our primary diagnostic tool for detecting discrepancies from this assumption is looking at **residuals**.
- Residuals are the distances between each y-value and the OLS line.

$$y_i - \hat{y}_i$$

- A useful diagnostic tool is the residual plot which plots the residuals against x or \hat{y} .
- Example: let's look at the residual plot of our life expectancy and national income data.

Residual scatterplot of life expectancy/national income data

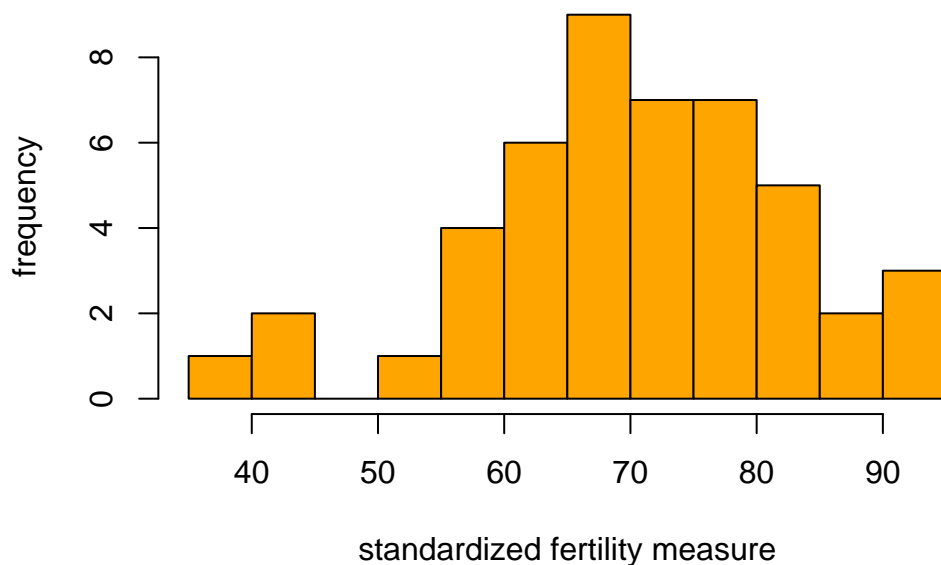


- * There should be no pattern to the residuals, and you should look for significant outliers.
- * Variation in residuals should not increase or decrease as x increases - show a cone: this indicates that our predictive ability depends on x.

2. Outliers and influential observations

- Like the mean and standard deviation, regression can be seriously affected by outliers.
- An outlier which significantly affects the slope and intercept of the line is referred to as an **influential point**. An influential point can be an outlier on either the independent or dependent variable.
- The first step is to examine the distributions of all variables used in the analysis to detect outliers and potential influential points. As an example, let's look at the distribution of fertility scores in our Swiss data:

Histogram of fertility index, Swiss cantons (1888)

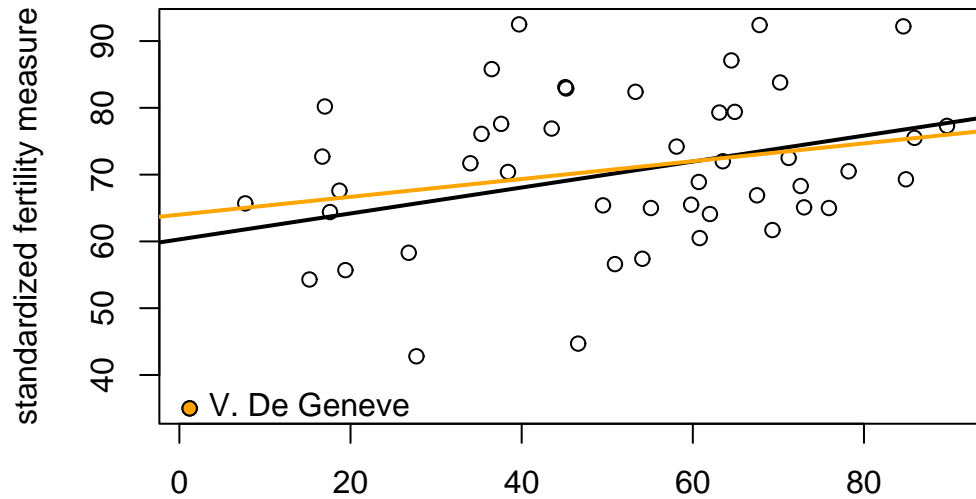


There are some potential outliers in the lower tail.

- Just because a data point is an outlier does not mean that it is an influential point. To be an influential point, the OLS regression line must be heavily affected by that

point. To show this, let's exclude the lowest value on the fertility measure from our swiss example.

Scatterplot of agriculture and fertility in Switzerland (1881)



% labor force in agriculture
based on data from 47 francophone Swiss cantons in 1888

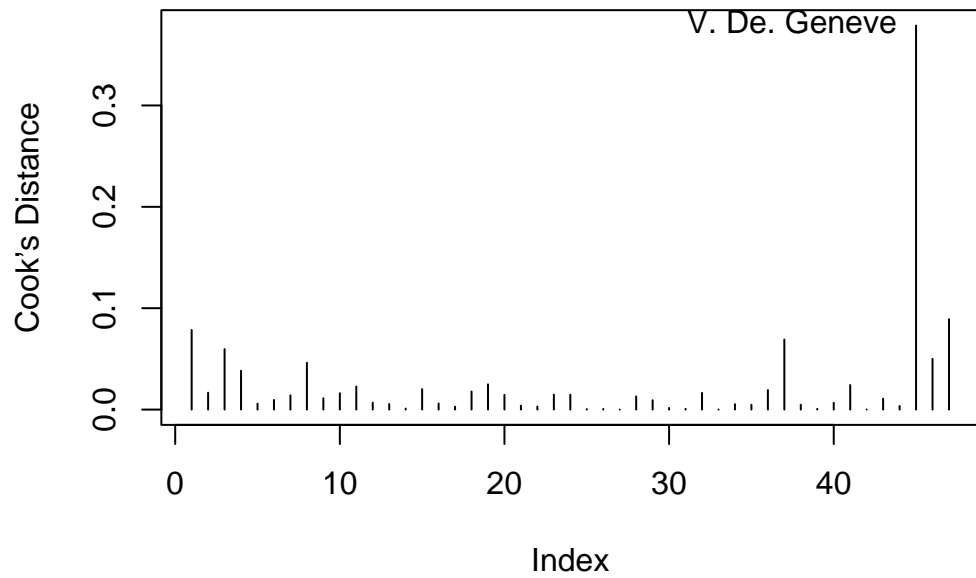
The change in the line here indicates the influence of the Geneva data point.

- It would be nice to be able to quantify this influence. We can do so using **cook's distance**. The formula for Cook's Distance is somewhat complex, but it gives a rough assessment of how large an effect the removal of a point has on the values of b .

$$D_i = \frac{\sum_j \hat{y}_j - \hat{y}_{j(i)}}{p * \sum_j (y_j - \hat{y}_j)}$$

There are formal ways to assess how high Cook's distance must be in order to be a concern. A less formal, rule-of-thumb is to be concerned about and Cook's distance greater than one. In our example, we can plot Cook's distance for each of our observations:

Cook's distance in Swiss data



The Geneva data point is clearly the most influential point, but its influence is not enough that we should really be concerned.

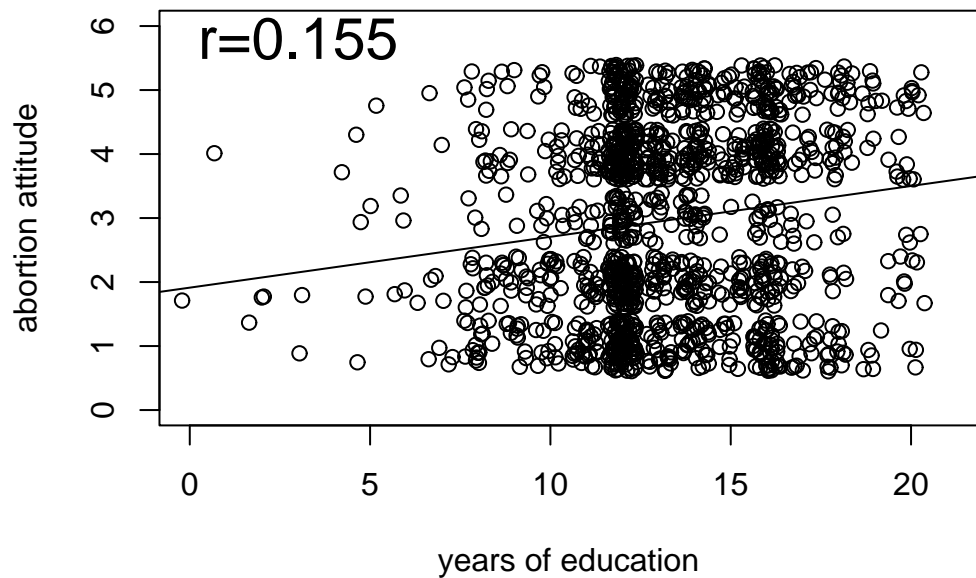
- Do not automatically remove outliers and influential points! Check the data, look for errors or explanations for the outlier. You should always have a substantive reason to remove outliers.

1.10 Regression diagnostics and cautions: aggregate data, collinearity, and causality

3. Problems with the use of **aggregate** data. Aggregating here means that the data are the means of some variable for the units of observations we care about across some higher level of units. For example, we might calculate the mean years of education in a state for the entire US.

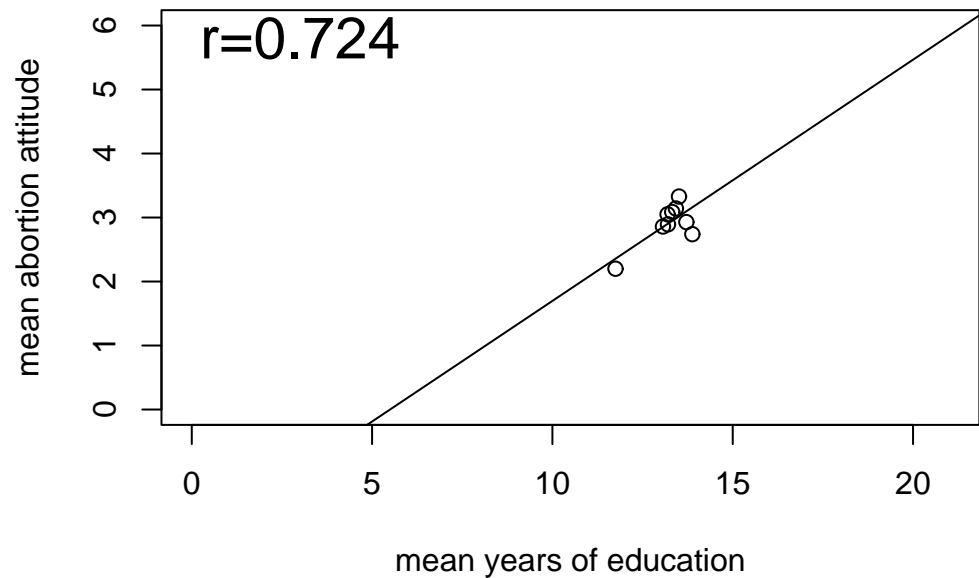
- Inflation of r in grouped data.
 - * Averaging smooths out variation at the individual level and thus leads to artificially high values of r . The r 's between grouped and individual data are not comparable.
 - * Show example with abortion attitudes and education across US regions.

scatterplot of education and abortion attitude, GSS 1994 with jitter



Explain the use of jitter. Now show the aggregate datapoints.

rplot of aggregate education and abortion attitude at the reg



The relationship between education and abortion attitudes looks far stronger at the aggregate level than it actually is at the individual level.

- A far more serious problem with aggregate data is the **ecological fallacy**.
 - * The ecological fallacy is the inappropriate assumption that relationships at the aggregate level will also hold at the individual level.
 - * It is entirely possible to find a relationship at the aggregate level that doesn't hold at the lower level.
 - * Example: William Robinson coined the term in 1950 when he looked at the relationship between literacy and the proportion of immigrants across states. he found a correlation of 0.53, indicating that areas with lots of immigrants were highly literate. However, at the individual level, the correlation between immigrant status and literacy was -0.11. Immigrants were less likely to be literate. The correlation at the aggregate level occurred because of a tendency for immigrants to settle in areas where the native-born population was highly literate (generally, urban areas).
- In some cases, one must think carefully about what the appropriate level is for

analysis. Modeling should occur at the level of the theorized process. (what about the Swiss Canton data, for example?)

4. Extrapolating beyond the range of data

5. Collinearity

- Two explanatory variables are **perfectly collinear** if knowing one automatically tells you the value of another. Examples
 - * Whether you are a man or a woman
 - * Percent of the week you spend working and the percent of the week you are not working
- You cannot include two perfectly collinear variables into a regression equation because they are both giving you the exact same information.
- Explanatory variables can also be approximately collinear if there is a high correlation between them (education and income, for example).
- This is problematic because it makes the coefficients for these variables harder to estimate. We will learn more about this problem later when we deal with problems of estimation.
- The quickest way to check for collinearity among your variables is to construct a **correlation matrix** of the variables and look for high r values. For our abortion data:

	Age	Education	Female
Age	1	-0.198	0.06
Education	-0.198	1	-0.017
Female	0.06	-0.017	1

6. **Correlation does not mean causation!!**

- Just because two variables are correlated does not mean they are causal. There is always the problem of lurking variables that may explain both variables (draw picture). This is sometimes called **spuriousness**. Examples:
 - (a) Taller children have higher standardized test scores. (taller children are older)
 - (b) Children with higher IQ scores get better grades. (Such children come from advantaged family backgrounds)
 - (c) Widows/Widowers have higher death rates than married people. (widows are older than married people on average)
- Well-designed experiments are the best way to avoid lurking variables, but these are rarely possible in the social sciences. So how do we establish causation:

- (a) Control for as many confounding variables as possible (we will learn how to do this later)
- (b) Establish temporal order
- (c) Be certain cause is plausible
- (d) Look for consistency across studies
- (e) **Tell a good story**
- (f) Be humble

1.11 Catch-up and review

2 Data Collection and Statistical Inference

2.1 Data collection (experiments and surveys)

- Introduction

- Where do our data come from and what are the limitations does this put on what we can say about the world?
- We are interested in the **generalizability** of our results.
- Types of data collection
 1. Anecdotal evidence: "My uncle once" - Haphazardly selected small number of observations - completely unreliable, not considered valid for scientific inquiry
 2. Observational study - We observe individuals (or process or units, etc) and (potentially) measure variables as they are in the population. A particularly important kind of observational study is one drawn from a **sample**, because if the sample is well chosen, then we can better generalize from the data. If we have a **census**, then we have entire population, and generalizability is not an issue.
 3. Experiment - We actively impose treatment on some individuals in order to observe a response. If we choose individuals for treatment well, then we will be able to make a causal connection between treatment and response.
- All of these methods have the potential to generate **bias** - that is to measure something in the population incorrectly.
 - * Random bias - bias due to random chance - we will learn how to measure our uncertainty here
 - * Systematic bias - bias due to technique of data collection or measurement

- Experiment

- Divide subjects **randomly** into at least one treatment group (**factors**) and a control group
- Treatment group(s) receive treatment
- Control group receives no treatment (problem of the placebo effect)
- Examine relationship between treatment and control group for some response variable (aspirin -> heart attack).
- Experiments provide good evidence for causation because treatment is completely randomly assigned and thus there are no lurking variables.

- Relationship still might be due to random chance, but as we will learn we can measure the likelihood of this.
- Cautions
 - * Placebo Effect
 - * Double-blind
 - * Producing realistic contexts (particularly a problem in social science research)
- More common in the social sciences are so-called ”**natural experiments.**” A natural experiment is a situation in which through some outside shock to a social system, a treatment effect is delivered which cannot be the cause of lurking variables. (example: Mariel boat lift)
- Survey Sampling Design
 - Interested in some characteristic of the population, so we sample from the population to measure this characteristic (i.e. Who would you vote for today in the presidential race?)
 - How is the sample drawn? (Is it likely to be biased?)
 1. Voluntary Response sample or **Self Selected Listener Opinion Poll** - Totally worthless, strong biases usually generated
 2. **Simple Random Sample** (SRS) - each sample of size n units from the population has an equally likely chance of being drawn. This is more complex than an equal chance for each individual. (divide the room in half and sample one member of each sub-group - this is not an SRS)
 3. **Stratified Random Sampling** - Divide population into strata along some dimension (income, race) and then draw SRS's within these strata.
 - * Oversampling - Often want to get more information about a small sub-population, so you will divide into strata and then **oversample** so the members of the over-sampled group have a better chance of entering sample.
 - * Often used with minority groups in the United States.
 - * When groups are oversampled, we use **weights** to make the overall sample representative again.
 - * **Population weights**, p_i are the inverse of the probability of the i th person being sampled. In an SRS 1% sample of the US, everyone would have a population weight of 100.

- * It is useful to scale the population weights so that they sum to the size of the sample. These are **sample weights** (w_i) and are calculated as:

$$w_i = \frac{p_i}{\bar{p}}$$

- * Weighted mean:

$$\bar{x} = \frac{\sum w_i x_i}{\sum w_i}$$

- * Population of 50 A individuals and 10 B individuals. We sample 5 A and 5 B. The prob for A is 1/10, while the prob for B is 1/2. The pop weights are 10 and 2 respectively, so the sampmle weights (mean of 1) are 10/6 and 2/6 (measuring children ever born)

Group	x_i	w_i	$w_i x_i$
A	7	10/6	11.667
A	4	10/6	6.667
A	8	10/6	13.333
A	11	10/6	18.333
A	5	10/6	8.333
B	2	2/6	0.667
B	1	2/6	0.333
B	0	2/6	0.000
B	3	2/6	1.000
B	4	2/6	1.333
			45
			10
			61.667
unweighted:			$\bar{x} = 45/10 = 4.5$
weighted:			$\bar{x} = 61.667/10 = 6.17$

4. **Cluser (Multistage) Sampling** - SRS of geographical units (cities, streets, etc.) then SRS's of individuals within selected geographical units. Often used to save money.

– Cautions

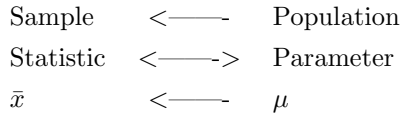
- * **Sampling frames** vs. populations (i.e. using a phonebook for presidential survey) (draw a Venn Diagram) - can cause undercoverage if sampling frame and population are significantly different.
- * **non-response** Can create systematic bias if those who decline to participate are systematically different than those who participate.

- * **Response bias** - Way question is phrased or characteristic of interviewer affects response.
 - Bad question: Do you oppose the murdering of unborn fetuses?
 - Black interviewer for white interviewee on questions about racial attitudes
- * Question might generate an opinion which was not pre-existent.

2.2 Sampling distributions and probability rules

- How can we generalize from our sample to the population? This is the fundamental question of statistical inference.

- **Parameter** - some measure in the actual population. Usually shown with a greek letter.
 μ =mean of some variable in the population
- **Statistic** - measurement in a sample of the population, usually intended to estimate a parameter. \bar{x} =mean of sample.



- **sampling variability** - Variability in reported statistic based on random chance. Therefore reported statistic will be somewhat different in different samples.
- What would happen if we took many samples?
 1. Take a large number of samples
 2. calculate statistic (\bar{x})
 3. make a histogram of \bar{x}
 4. Look at the shape of this histogram
- This is an approximation of the **sampling distribution** - The distribution of values taken by a statistic in all possible samples of the same size from the same population.
- A statistic is **unbiased** if the center of its sampling distribution is the true value of the parameter in the population.
- Variability of a statistic (random bias) - largely determined by sample size n . As n goes up, statistic variability goes down. (the spread of the sampling distribution shrinks).
- (Draw a picture)
- A statistic is **efficient** if its sampling variability is smaller than another statistic (pick the first value of sample as an estimator of μ).
- The **most efficient** statistic is the statistic with the smallest sampling variability among all possible statistics. What we want is the minimum variance unbiased estimator of our parameter, or BUE (best unbiased estimator).

- The idea of probability
 - Conduct independent trials of some process (say flipping a coin)
 - **Over the long run**, a certain regularity will appear
- Basics of probability
 - **Sample space** S - set of all possible outcomes
 - one coin toss, $S = \{H, T\}$ or $S = \{0, 1\}$
 - two coin toss, $S = \{HH, TT, HT, TH\}$ or $S = \{0, 1, 2\}$
 - continuous range of 0 to 1, $S = [0, 1)$
 - **Event** - outcome of random phenomenon, in other words a subset of the sample space
 - A = event - say a certain sequence of heads and tails
 - Every event has a **probability** associated with it, $P(A)$
 - * Probability is between 0 and 1, $0 \leq P(A) \leq 1$
 - * $P(S) = 1$
 - * $P(A^c) = 1 - P(A)$
 - * Draw a venn diagram
 - * An **intersection** of two events means that they both happen.

$$P(A \cap B)$$

- * A **union** of two events is the probability that either one of them occurs.

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

- * **Disjoint events** have an intersection of 0. In this case,

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

- * **Conditional probabilities** The probability of event A, given that we observe event B.

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

- * **Independence of events** Knowing whether event B happened or not, does not tell you anything about the probability of event A happening.

$$P(A|B) = P(A)$$

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

– Go through example with deck of cards

1. What is the probability of a face card?

A=a face card

$$P(A) = \frac{12}{52} = 0.23$$

What is the probability of a non-face card?

$$P(A^c) = 1 - P(A) = 1 - .23 = .77$$

What is the probability of a diamond suit?

$$P(B) = \frac{13}{52} = .25$$

2. What is the probability of getting a king given that we got a face card?

$$P(A \cap C) = \frac{4}{52} = .0769$$

$$P(C|A) = .0769/.23 = 0.33$$

What is the probability of getting a king given that we got a diamonds?

$$P(C \cap B) = \frac{1}{52} = .019$$

$$P(C|B) = .019/.25 = .0769$$

Is D independent of A?

3. What is the probability of getting a face card or an ace?

$$P(A \cup E) = \frac{12}{52} + \frac{4}{52} = \frac{16}{52} = .308$$

Sine events are disjoint, we do not need to subtract intersection.

4. What is the probability of getting a face card or a card higher than 8

$$P(A \cup F) = \frac{12}{52} + \frac{20}{52} - \frac{12}{52} = \frac{20}{52} = .385$$

5. You draw three cards with replacement. What is the probability of getting at least one face card.

Use the complement. Are the draws independent?

$$P(G^c) = \left(\frac{40}{52}\right) * \left(\frac{40}{52}\right) * \left(\frac{40}{52}\right) = 0.455$$

2.3 Random variables and their probability distributions

- Give the basic intuition behind the integral from calculus - the sum of the area under $f(x)$ from a to b .
- If event can be recorded numerically (number of heads on n coin tosses), we refer to it as a **random variable**. Usually represented as a capital letter from the end of the alphabet.
 X =number of heads in 2 coin tosses
- A random variable can be discrete (takes certain exact values like whole numbers) or continuous (has some probability of taking all possible numbers in some range).

- Every random variable has a **probability distribution** associated with it which gives:
 For discrete random variables:

$$P(X = x) = p(x)$$

For continuous random variables:

$$P(a \leq X \leq b) = \int_a^b f(x)dx$$

- Example, X =number of heads on 2 coin tosses

Outcome	x	$P(X = x)$
TT	0	$1/2 * 1/2 = 1/4$
HT or TH	1	$1/2 * 1/2 + 1/2 * 1/2 = 1/2$
HH	2	$1/2 * 1/2 = 1/4$

- Because probability distributions cover the entire sample space, they must add up to 1.
- The normal density curve is an example of a continuous distribution
- Mean and variance of probability distributions
 - Every random variable has a mean and variance based on its probability distribution..
 - The mean of a random variable is called its expected value (EX). For discrete variables, EX is given by:

$$EX = \sum x * p(x)$$

- The calculation of EX for continuous distributions is similar but requires higher mathematics:

$$EX = \int_{-\infty}^{\infty} x * f(x)dx$$

- The variance of X is given by VX.

$$VX = E(X - EX)^2 = EX^2 - (EX)^2$$

Outcome	x	$p(x)$	$x * p(x)$	$x^2 p(x)$
TT	0	.25	0	0
HT or TH	1	.5	.5	.5
HH	2	.25	.5	1

$$EX = 0 + .5 + .5 = 1$$

$$VX = EX^2 - (EX)^2 = (0 + .5 + 1) - 1^2 = 1.5 - 1 = 0.5$$

- When we have two random variables, we can speak of the covariance between them - analogous to our correlation coefficient

$$Cov(X, Y) = E[(X - EX)(Y - EY)]$$

If we divide by the square root of variances of X and Y, we get ρ which is analogous to the correlation coefficient.

- Addition rules (add two separate tosses of 2 into a single toss of 4)

$$E(X + Y) = EX + EY$$

$$V(X + Y) = VX + VY + 2 * Cov(X, Y)$$

- There are several theoretical probability distributions for certain kinds of random variables. Two of these distributions will be important for us.

1. The **Binomial** Distribution

- Used when we have a binary outcome (true/false, yes/no, heads/tails, etc.)
- X is a random variable given the number of “successes” in n trials where the probability of success on any given trial is p . The trials are independent.

$$X \sim \text{bin}(n, p)$$

- The probability distribution of X is given by:

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

(go through the formula step-by-step)

- $\binom{n}{k}$ tells us how many different ways we can combine n successes and $n - k$ failures. (do example of 2 true and 1 false).
- $p^k(1 - p)^{n-k}$ is the probability that we get k successes and $n - k$ failures in any order.
- The expected value and variance of X are given by:

$$EX = np$$

$$VX = npq$$

- The **Bernoulli** distribution is a special case of the binomial distribution where $n = 1$.

2. The **Gaussian** (Normal) Distribution

- This is the classic bell-shaped distribution. It approximately describes the distribution of many things, but its importance is chiefly tied to the central limit theorem which will discuss tomorrow.

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

- Don't worry, you don't need to know this, but what you should notice is that differences in the shape of this curve are completely determined by μ and σ . (show example of different distributions).
- μ provides a measure of center and $EX = \mu$
- σ provides a measure of spread and $VX = \sigma^2$.
- We can use a shorthand to describe a particular normal curve: $N(\mu, \sigma)$
- 68%, 95%, 99.7% rule: For any normal curve, 68% of the area is within one σ of the mean, 95% is within two σ 's of the mean, and 99.7% within three σ 's of the mean. (Draw a picture)
- A useful trick that will serve us well is to standardize any normal distribution to make it easily comparable. We want to turn a $N(\mu, \sigma)$ distribution into a $N(0, 1)$ distribution. We can make any particular x from a normal curve comparable by computing a **z-score**.

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

- When we subtract μ we are re-centering the distribution around zero (draw a picture which shows the distribution shifted).
- When we divide by σ , we are changing our unit of measurement to standard deviations from the center.

- Example: re-scale one of Babe Ruth's homeruns (60)

$$\bar{x} = 41.12, s = 13.56$$

$$z = \frac{60 - \bar{x}}{s} = \frac{60 - 41.12}{13.56} = 1.39$$

- Show Table A, go through examples:

(a) $z < 0$ (.5)

(b) $-0.5 < z < 1$ (.8413-.3085=.5328)

(c) $z > 1.96$ (1-.9750=.025)

2.4 The sampling distribution

- Population Distribution vs. Sampling Distribution

Population Distribution Distribution of some variable in some population. It is also the probability distribution of a random variable which is the value of one draw from the population.

Sampling Distribution If you take a random sample from a population, a measured statistic (\bar{x}) is a random variable. The random variable of a statistic has a probability distribution. i.e. What is the probability that the mean of our sample is between a and b?

- What are the sampling distributions of means and sums?
 - If we take a random sample of size n from a population on some characteristic. As an example, let's treat this class as a population and look at the characteristic of height. I could take some sample of the class of size n .
 - The population mean and variance of this characteristic are μ and σ^2 , respectively. (show the values and distribution for the class)
 - we essentially have n **independent identically distributed** random variables. The mean and sum are just simple functions of these random variables:
 X_i i.i.d with some $E(X) = \mu$ and $V(X) = \sigma^2$, but could be any distribution

$$\bar{x} = \frac{1}{n}(X_1 + X_2 + \dots + X_n)$$

$$Y = X_1 + X_2 + \dots + X_n$$

- What is the expected value of \bar{x} ($\mu_{\bar{x}}$)?

$$\begin{aligned} E(\bar{x}) &= E\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] & (1) \\ &= \frac{1}{n}(E(X_1) + E(X_2) + \dots + E(X_n)) \\ &= \frac{1}{n}(\mu + \mu + \dots + \mu) \\ &= \frac{n\mu}{n} \\ &= \mu \end{aligned}$$

By similar logic,

$$E(Y) = n\mu$$

- What is the variance of \bar{x} ($\sigma_{\bar{x}}^2$)?

Because random variables are independent we can add them.

$$\begin{aligned}
 V(\bar{x}) &= V\left[\frac{1}{n}(X_1 + X_2 + \dots + X_n)\right] & (2) \\
 &= \frac{1}{n^2}(V(X_1) + V(X_2) + \dots + V(X_n)) \\
 &= \frac{1}{n^2}(\sigma^2 + \sigma^2 + \dots + \sigma^2) \\
 &= \frac{n\sigma^2}{n^2} \\
 &= \frac{\sigma^2}{n}
 \end{aligned}$$

By similar logic,

$$V(Y) = n\sigma^2$$

- So SD of mean is σ/\sqrt{n} and SD of sum is $\sigma\sqrt{n}$. Both of these imply that the SD is shrinking as n increases (in the latter case relative to the size of the mean). This is a nice property. In a random sample, our estimate of the mean of the population will be unbiased and the variation in our estimate will decline as we draw larger and larger samples.
- Example: show the distribution of the mean for all possible class samples for $n = 4, 8, 10$.
- What about the shape of this distribution?

- If our underlying population distribution is normal (i.e. $X_i \sim N(\mu, \sigma)$), then our sampling distribution will also be normal.
- lots of population distributions are not normally distributed so this is of limited use. But as n increases we will be able to make a generalization about all sampling distributions.
- **The Central Limit Theorem** - as n gets large, the sampling distribution of means and sums becomes normal no matter what underlying population distribution they are drawn from.

$$\bar{x} \sim N(\mu, \sigma/\sqrt{n})$$

$$Y \sim N(n\mu, \sigma\sqrt{n})$$

- The further the underlying population distribution is away from the mean, then the larger n will have to be. In many situations, however, 100 to 200 cases will be enough.

2.5 Sampling distribution, the binomial case

- Let's take an example of a population distribution.
 - The population is divided into two categories (yes/no, agree/disagree, black/white). For our purposes we will call the division "successes" and "failures." The population distribution has one important parameter, p , which gives the proportion of successes in the population. (draw a bargraph to capture this population distribution).
 - We are going to take a sample from this population of size n . Each draw from the population will be independent and they will all come from the same population, so they will be identically distributed.
 - We will "count" the number of successes in each draw. So each observation is a random variable X_i which takes a 1 if we get a success and a 0 if we get a failure.
 - There is a name for such a distribution. Each X_i is distributed as a **bernoulli** random variable with parameter p .
 - Let's say we want the count of successes in our sample of n (Y) and we also want the proportion of successes in our sample ($X/n = \hat{p}$).
 - The counts of success is simply the sum of n bernoulli random variables and the proportion is simply the mean of n bernoulli random variables.
- What are the sampling distributions for our count and proportion?
 - We have already done this with our probability of getting no face cards (no successes) in four draws from a deck of cards.
 - The probability of observing k successes in n draws is given by the binomial distribution (give the formula and go over it with simple example of rolling two sixes).
 - The proportion is not distributed like a binomial, but you can go from proportions to counts to get the right distribution.
 - It turns out that $E(Y) = np$ and $V(Y) = npq$, $E(\hat{p}) = p$ and $V(\hat{p}) = pq/n$.
- But the counts of success is simply the sum of n bernoulli random variables and the proportion is simply the mean of n bernoulli random variables, so if n is large enough we can just use our normal approximation rather than the exact binomial formula.

$$\hat{p} \sim N(p, \sqrt{pq/n})$$

$$Y \sim N(np, \sqrt{npq})$$

- Binomial is discrete, but normal is continuous. So in order to get the best approximation, we need to use the continuity correction.

$$P(Y = 9) \rightarrow P(8.5 \leq Y \leq 9.5)$$

$$P(Y \leq 9) \rightarrow P(\leq 9.5)$$

And so forth.

- In the population of registered voters, Candidate A is preferred to Candidate B 53% to 47%. We are a poor pollster and so we draw a sample of 100 registered voters. What is the probability that our sample shows Candidate B beating Candidate A? What if we drew 1000 registered voters?

2.6 Confidence intervals and hypothesis tests

- Confidence Intervals

- Our estimate of the population mean based on the sample mean is often called a **point estimate**, because it gives a precise value.
- We often report not just our point estimate, but our confidence in the range of possible values for the true population mean. This range is called a **confidence interval**.
- If we have a large enough sample that the CLT holds, then we can actually calculate a confidence interval.
- The logic is as follows:
 - * Probability of C% that \bar{x} will be within $z\sigma_{\bar{x}}$ of μ (draw picture).
 - * Reverse this: μ will be within $z\sigma_{\bar{x}}$ of \bar{x} .
 - * So C% of all samples will contain μ in the interval between $\bar{x} \pm z\sigma_{\bar{x}}$.

C	z
68	1
95	1.96
99.7	3

- Draw a picture of 20 CI's from 20 samples around true population mean.
- Confidence is not exactly the same thing as probability - 95% of the time we are right.
- Confidence intervals are what is often called "margins of error."
- Example: Zogby Poll 11/1. Bush=.502, Kerry=.498, N=955

$$s_{\bar{x}} = \sqrt{.502 * .498 / 955} = .016$$

$$\bar{x} \pm 1.96s_{\bar{x}} = .502 \pm .031 = (.471, .533)$$

- How do you shrink confidence intervals?
 - * Increase sample size
 - * reduce σ
 - * Use lower level of confidence

- Hypothesis Testing

- Usually we are not just interested in uncertainty, but rather in testing some specific claim or hypothesis.

- Such tests are called hypothesis tests.
- Hypothesis tests begin with two hypotheses:
 - * The **null hypothesis** - Generally this is usually set up as something we want to argue against and as an exact equality. Using the poll conducted by zogby, our null hypothesis might be that the presidential race was tied.

$$H_o : p = .5$$

- * The **alternative hypothesis** - Generally, this is the hypothesis which we are interested in. There are two important variants of the alternative hypothesis:
 - **Two-sided test** - In a two-sided test, we assume the difference from the null hypothesis could go in either direction.

$$H_a : p \neq .5$$

- **One-sided test** - We assume for a-priori reasons that the effect could only be in one direction.

$$H_a : p > .5$$

- Based on the data from our observed sample, we will either reject or not reject the null hypothesis.
- We never accept or do not accept the alternative hypothesis.
- We assess the two hypotheses by calculating a p-value. A p-value is the probability that we would get a sample mean of \bar{x} or larger conditional on the null hypothesis being true. (draw picture, show difference between two-sided and one-sided)
- Let's try this for the sample problem. Assume normal distribution because n is so large.

$$z = \frac{.502 - .5}{.016} = .13$$

The probability to the right of this z is $1 - .5517 = .4483$. For a two-sided test, we would double this p-value to 0.8966. In this case, the z -score is often called a **test statistic**.

Definition: On a sample of size 955, if the race was truly tied, we would expect to get a value of .502 or higher, 44.83% of the time.

Clearly, this is not a low enough value that we would feel comfortable rejecting the null hypothesis.

- Fixed α . Traditionally, researchers preset critical p-values at which they will reject the null hypothesis. The most common is the 5% level, but 10% and 1% are not uncommon. These values are generally agreed upon as reasonable, but are also purely arbitrary.

- When a p-value is below α , the result is reported as statistically significant. A much better term would be "statistically distinguishable from ..." where ... is some statement of the null hypothesis. In our polling case our results are "statistically indistinguishable from a 50/50 tie."

2.7 Errors in hypothesis testing: statistical significance and power

- Two kinds of mistakes can be made with hypothesis testing:

	H_0 is true	H_a is true
H_0 not rejected	Correct	Type II Error
H_0 rejected	Type I error	Correct

- Statistical significance is focused on minimizing Type I errors - i.e. Under the assumption that the null hypothesis is true, what is the probability of accidentally rejecting it. Weight is given to null hypothesis, burden of proof is on alternative hypothesis. (false positive)
- What about a Type II error - this is the probability of not rejecting the null hypothesis when it is false, i.e. the probability of not observing a difference when there really is one. (false negative) We made a type II error with our presidential prediction.
- These two errors are related - for a given sample size, decreasing the probability of making one error will increase the probability of making the other.
- Both of these errors can be reduced by increasing sample size.
- Example with statistical significance. Increase Zogby poll to 9950 individuals.

$$s = \sqrt{.502 * .498 / 9950} = .005$$

$$z = .02 / .005 = 4$$

There is a very small probability that Bush's lead was by chance if our poll was 10 times as big.

- Statistical Power Calculations
 - Statistical power is the probability that the test will detect a difference if one truly exists. (probability of not making a type II error when H_a is true)
 - Trickier calculation because we have to set an α rejection level and presume the level of the true difference.
 - Example: Bush's support was really 51.5%. What was the power of Zogby's poll to detect this true lead of 1.5% points? Assume we will reject at $\alpha = .05$

1. For a one-sided test, what z-score do we need to reject at .05 level. Roughly $z = 1.645$.

What does this correspond to in real terms?

$$\sigma = \sqrt{.515 * .485 / 995} = .0158$$

$$\bar{x} = 1.645 * .0158 + .5 = 0.526$$

2. When the true value of p is .515, we would require a sample mean of .526 for a sample of 995 people to reject the null hypothesis that the race is a tie. What is the probability that we will observe a value this high or higher?

$$\frac{.526 - .515}{.0158} = 0.69$$

The probability of $z > 0.69$ is 0.2451. That is the statistical power of the poll to detect that Bush was going to be the true winner. Not very high.

- Two-step process to find statistical power
 1. Find the value of \bar{x} required to reject the null hypothesis.
 2. Find the probability of observing a \bar{x} of that size or (larger,smaller) under some assumption about the true mean.
- Most good studies should have a statistical power of 80% or so, but you never know exact statistical power because you don't know how large the potential difference is between H_0 and H_a .
- Cautions about hypothesis testing
 1. You never can accept or reject the alternative.
 2. Balancing statistical significance and statistical power
 3. Arbitrariness of α What if p-value is .051?
 4. substantive vs. statistical significance
 5. Inference not valid for all data sets
 6. Statistical insignificance can be interesting
 7. If you run a bunch of hypothesis tests, 1 in 20 will be stat sig by random chance

2.8 t-tests and the t-distribution

- Problem with previous examples

- I played a trick on you: We don't usually know the population standard deviation σ .

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

- Therefore, we can't estimate the z-scores we used in our hypothesis test.
- Normally, we replace our σ with our sample estimate s .
- This is problematic because s itself is an estimate and thus there will be greater uncertainty in our test.
- In fact when we replace σ with s , our test statistic will no longer have the standard normal distribution (**Draw a picture to remind them**), but rather come from what is known as the **t-distribution**.

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

- The t-distribution

- A t-distribution has one parameter: degrees of freedom, which is given by $n - 1$.
- A t-distribution looks similar to the standard normal distribution except that it has fatter tails and a smaller peak. This difference reflects the higher degree of uncertainty.
- As the degrees of freedom increase (the sample size increases), our t-distribution looks more and more like the standard normal.
- (Pass out a copy of the table from the back of the book) this table gives the critical values for upper tail probabilities. You see that the necessary critical values shrink as df increases. When $df = 1000$, the critical values for the t-distribution are very close to the critical z-values from a standard normal table. (show the example of $.025 - \alpha = .05$ -, where $z = 1.96$).

df	t
20	2.086
50	2.009
100	1.984
1000	1.962

- Using t

- We can substitute t in our equation for confidence intervals.

Before:

$$\bar{x} \pm z(\sigma/\sqrt{n})$$

Now:

$$\bar{x} \pm t(\sigma/\sqrt{n})$$

With 20 degrees of freedom, we would use $t = 2.086$ for a 95% confidence interval.

- We can use the t-distribution for our one-sample hypothesis tests, where $df = n - 1$.
- For small n , this will still not be accurate if underlying population distribution is non-normal.
- However, it turns out that the t-distribution is fairly **robust** to violations of the normality assumptions, meaning it gives a good approximation even for small sample sizes. Use the following rule of thumb:
 - * $n < 15$, if data look close to normal with no outliers, use t-test
 - * $15 \leq n < 40$, use t-test except in cases of big outliers or extreme skewness
 - * $n \geq 40$, use t-test and you should be approximately accurate.

- The two-sample t-test

- So far we have compared one sample to a hypothesized population mean. More often what we are interested in is comparing the means of two groups for which we only have samples.
- We call this the **two-sample t-test**.
 - μ_1 =mean for the first group
 - μ_2 =mean for the second group
- We are interested in $\mu_1 - \mu_2$, i.e. the difference between the groups.
- $H_0 : \mu_1 - \mu_2 = 0$, means are the same
- $H_a : \mu_1 - \mu_2 \neq 0$, two-sided alternative
- $H_a : \mu_1 - \mu_2 > 0$, a one-sided alternative
- Our observed difference is $\bar{x}_1 - \bar{x}_2$.

$$E(\bar{x}_1 - \bar{x}_2) = \mu_1 - \mu_2$$

$$V(\bar{x}_1 - \bar{x}_2) = \frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

- Our test statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- We are estimating **two** standard deviations here, so it turns out that our test statistic is only approximately distributed as a t . Fairly robust, particularly, if groups are equal in size.
- What should the degrees of freedom be? Two options:

- * Take the smaller of $n_1 - 1$ or $n_2 - 1$. This will give you a conservative estimate on the degrees of freedom.
- * Use a complicated formula, which has little intuition.

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

- * This formula gives very precise results for $n > 5$. It is always as large as the smaller sample size (minus one) and never larger than the largest sample size (minus one).
- If you assume that $\sigma_1 = \sigma_2$, then this equation simplifies somewhat.

- * Calculate a pooled estimate of the variance - just a weighted average

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

- * Then the variance of the difference between means simplifies

$$\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}$$

$$\sigma^2 \left(\frac{1}{n_1} + \frac{1}{n_2} \right)$$

- * So the t-statistic is:

$$t = \frac{(\bar{x}_1 - \bar{x}_2)}{s_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$$

- * The degrees of freedom are $n_1 + n_2 - 2$
- * When we work with proportions, the pooled variance rule always holds. When we work with other continuous variables, however, it is an assumption.

- Example: let's take the observed age difference between survivors and non-survivors on the Titanic. Earlier in the semester, we calculated that survivors of the Titanic were 1.63 years

older on average than non-survivors. Taking this particular sinking as one of many possible sinkings, is this difference due to random chance or is there a real age advantage?

$$\bar{x}_s = 30.54$$

$$\bar{x}_d = 28.92$$

$$n_s = 427$$

$$n_d = 619$$

$$s_s = 15.06$$

$$s_d = 13.92$$

– Calculate numerator of test statistic:

$$\bar{x}_f - \bar{x}_m = 30.54 - 28.92 = 1.62$$

– Calculate denominator of test statistic:

$$\begin{aligned} & \sqrt{\frac{15.06^2}{427} + \frac{13.92^2}{619}} \\ & \sqrt{\frac{226.80}{427} + \frac{193.77}{619}} \\ & \sqrt{0.53 + 0.31} \\ & \sqrt{0.84} \\ & 0.92 \end{aligned}$$

– Calculate t

$$t = 1.62/0.92 = 1.76$$

– Let's use conservative estimate of degrees of freedom and a two-sided test. What is the critical value for $\alpha = .05$?

– Now let's pool our estimate of the standard deviation:

$$\begin{aligned} & \sqrt{\frac{427 * 226.80 + 619 * 193.77}{427 + 619}} \\ & \sqrt{\frac{216787.20}{1046}} \\ & \sqrt{207.25} \\ & 14.40 \end{aligned}$$

– So our test statistic becomes:

$$\frac{1.62}{14.40\sqrt{\left(\frac{1}{427} + \frac{1}{619}\right)}}$$
$$\frac{1.62}{0.91}$$
$$1.78$$

2.9 Contingency tables and Chi-squared tests

- We have now tested for differences between means of groups, but what about testing for differences on two categorical variables?
- Example: Let's use the relationship between class and survivorship in the Titanic data once again.

	Survived	Died	Total
1st Class	200	123	323
2nd Class	119	158	277
3rd Class	181	528	709
Total	500	809	1309

- Let's review some terminology.
 - f_{ij} refers to the count of observations in the i th row and the j th column.
 - **Joint distribution** - the share of each cell to the total

$$\frac{f_{ij}}{N}$$

where N is the total size (in our example, $N = 1309$).

- **Marginal distribution** - the distributions of each variable regardless of the other one

$$\frac{f_{i.}}{N} = \frac{\sum_{j=1}^J f_{ij}}{N}$$

$$\frac{f_{.j}}{N} = \frac{\sum_{i=1}^I f_{ij}}{N}$$

In other words, sum over the columns and the rows.

Put it all together and we have:

	Survived	Died	Total
1st Class	200 (0.094)	123 (0.153)	323 (0.247)
2nd Class	119 (0.121)	158 (0.091)	277 (0.212)
3rd Class	181 (0.403)	528 (0.138)	709 (0.542)
Total	500 (0.381)	809 (0.618)	1309 (1.00)

- **Conditional distributions** - the distribution of one variable conditional on being at a certain value on the other one. We saw before that the distribution of survivorship was very different given class in our sample.

	Survived	Died	Total
1st Class	200/323=0.62	123/323=0.38	1
2nd Class	119/277=0.43	158/277=0.57	1
3rd Class	181/709=0.26	528/709=0.74	1

- Notice these conditional distributions should sum up to one across the rows but not the columns. It should be clear that knowing the race of the wife changes the distribution of husband's race considerably.
- We want to test whether these two variables are related or not. The two variables are unrelated if knowing the value of one doesn't alter the distribution of the other. So how can we tell the relationship we observe in the conditional distribution is not due to sampling variability?
- Hypothesis test
 H_0 : row and column variable are not related
 H_a : row and column variable are related
- We run a chi-squared (χ^2) test.

1. Calculate expected cell frequencies (F_{ij}) under the assumption of independence.

$$F_{ij} = N * P(i \text{ and } j) = N * p(i)p(j) = N * \frac{f_{i.}}{N} \frac{f_{.j}}{N} = \frac{f_{i.} f_{.j}}{N}$$

	Survived	Died
1st Class	323*500/1309=123.4	323*809/1309=199.6
2nd Class	277*500/1309=105.8	277*809/1309=171.2
3rd Class	709*500/1309=207.8	709*809/1309=438.2

2. We need a measure of how far off our numbers are from those calculated under the assumption of independence. Two possibilities:

$$\chi^2 = \sum_i \sum_j \frac{(f_{ij} - F_{ij})^2}{F_{ij}}$$

$$L^2 = 2 \sum_i \sum_j f_{ij} \log(f_{ij}/F_{ij})$$

The first one is fairly intuitive. It gives the squared difference relative to the expected cell size.

3. When the two variables are independent, we would expect both of these test statistics to be zero, but because they are drawn from a random sample, they will be random variables. Thus, we need to know their sampling distribution in order to complete the test. It turns out that if H_0 is true, and N is large enough, both of these test statistics will be drawn from a χ^2 distribution with degrees of freedom equal to:

$$DF = (I - 1)(J - 1)$$

Where I and J are the number of rows and columns respectively.

4. What is the probability of getting a value greater than or equal to the test statistic in our sample on this distribution?

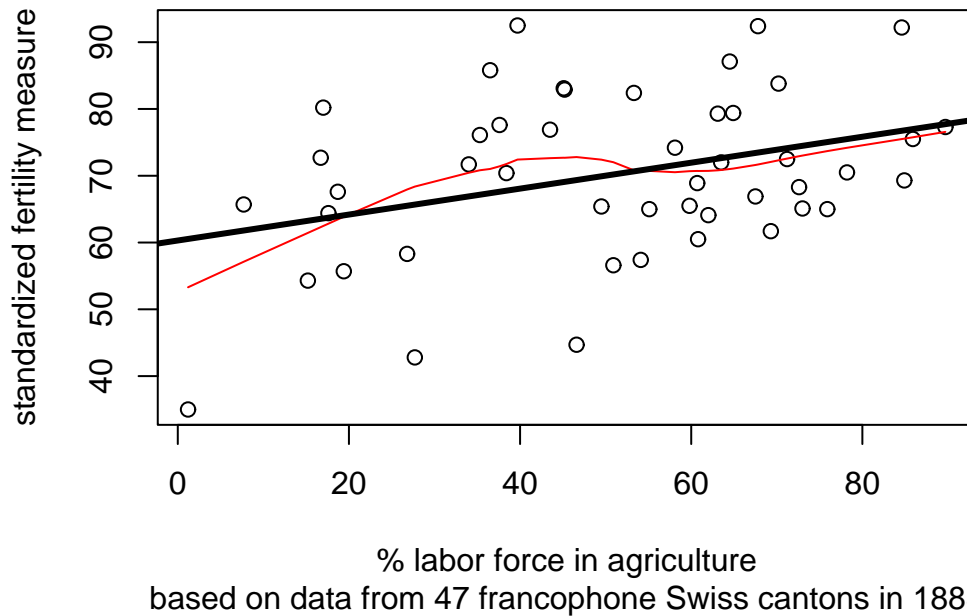
In our particular case, the χ^2 value is 127.86. We have 2 degrees of freedom. From the distribution at the back of the book, in order to reject at the 5% level, we need a χ^2 value of 5.99 or greater. Clearly, the relationship between class and survivorship in our sample is not due to random chance.

- In the particular case of the 2x2 table, we could also compare the conditional distributions using a two-sample t-test for proportions. We would get exactly the same result, but this technique is not generalizable to larger tables like the one we have here.

2.10 Inference for bivariate regression

- Review OLS regression
 - Plot scatterplot of swiss canton data:

Scatterplot of agriculture and fertility in Switzerland (1881) with OLS regression line and lowess smoother ($f=.66$)



- OLS regression line is the best-fitting line through these points. How do we decide what is "best?" We minimize the sum of the squared residuals.
- Regression line as we know it has two parts (write equation):
 - * Intercept - value of y when x is zero.
 - * Slope - change in y resulting from a unit change in x .
- The Population Regression Line
 - If we had the recorded values of y and x for an entire population, we could draw a scatterplot and regression line for this population. This would be the population regression line.

- We write the equation for it slightly differently, because we think of its slope and intercept as parameters.

$$\mu_{y|x} = \beta_0 + \beta_1 x$$

- The value $\mu_{y|x}$ is the mean of y conditional on some value of x . The equation tells us the expected value of y when we know x .
- There is some variability around this expected value of y . We assume that this variability is constant, regardless of the particular value of x :

$$V(y|x) = \sigma^2$$

- If y and x are independent of one another, what does that imply for our equation?

$$\beta_1 = 0$$

$$\mu_{y|x} = \beta_0 = \mu_y$$

- Simple Linear Regression Model

- At the individual level our linear regression model can be written another way.

$$y_i = \beta_0 + \beta_1 x_i + \epsilon_i$$

- This equation is made up of three logic parts:
OBSERVED=STRUCTURAL+STOCHASTIC
where,
OBSERVED -> y_i
STRUCTURAL -> $\hat{y}_i = \beta_0 + \beta_1 x_i$
STOCHASTIC -> ϵ_i

- There are different ways to view the above schematic:
Causation: observed=true process+disturbance
Prediction: observed=predicted+error
Description: observed=summary+residual
- Formula is always the same, but way of thinking about what you are doing is not - we will generally prefer the last method as it involves the least amount of baggage.
- The residuals can be thought of as random terms after the structural aspect of the model are fit. We can then ask what distribution(s) are these residual terms drawn from?

- * We assume that these ϵ_i terms are all drawn from the **same** distribution and that they are drawn **independently** of one another.
- * We call this assumption **i.i.d.** (Draw the scatterplot in reverse to show how we are thinking of the process)
- * We don't make any assumptions about the shape of this distribution, but we do assume that it is centered on 0 and has a standard deviation of σ .

– Our regression model has three parameters:

- * The intercept (β_0)
- * The slope (β_1)
- * The standard deviation of residuals around the line (σ)

• Estimating the parameters

– Normally, we only have a sample and so we are forced to estimate the parameters from our sample - thus, we will have to use the techniques of statistical inference.

Parameter	Statistic
β_0	b_0
β_1	b_1
σ	s

– Our regression equation for the sample is:

$$y_i = b_0 + b_1x_i + e_i$$

– We already have estimators for β_0 and β_1 :

$$b_1 = r \frac{s_y}{s_x}$$

$$b_0 = \bar{y} - b_1\bar{x}$$

– These estimator are unbiased. What does that mean?

– s is just the standard deviation of the observed error term e_i

$$e_i = y_i - \hat{y}_i$$

$$s^2 = \frac{\sum e_i^2}{n-2} = \frac{\sum (y_i - \hat{y}_i)^2}{n-2}$$

$$s = \sqrt{\frac{\sum (y_i - \hat{y}_i)^2}{n-2}}$$

- Go through this in steps. We take two off of n because two estimates were used to construct \hat{y}_i .

- Testing Hypotheses

- The most common hypothesis to test is whether there is a relationship between x and y or whether the observed relationship in the sample is a result of random chance.

$$H_0 : \beta_1 = 0$$

$$H_a : \beta_1 \neq 0$$

- Our test statistic will be:

$$\frac{b_1 - 0}{s_{b_1}}$$

- Where s_{b_1} is the estimated standard deviation for the sampling distribution of b_1 (NOT the same as s)
- What do you think the distribution of this test statistic might be? If N is large enough, it is a t-distribution. Even for small N , using the t-distribution for this statistic will give answers that are often reasonably close, but watch out for outliers and skewness in either x or y .
- Sometimes, researchers additionally assume that the error terms are drawn from a $N(0, \sigma)$ distribution. With this additional assumption, the t-statistic is valid for all sample sizes (but you still need a t because you have estimated s_{b_1}). This assumption can be checked with a normal quantile plot of the residuals - Watch out for outliers in small samples. If this assumption does not hold, your estimates of β_0 and β_1 will still be unbiased but your estimate of σ will be too small (inefficient).
- How do we get s_{b_1} ? Well the derivation is a bit complex, but here is the end result:

$$s_{b_1} = \frac{s}{\sqrt{\sum(x_i - \bar{x})^2}}$$

$$s_{b_0} = s \sqrt{\frac{1}{n} + \frac{\bar{x}^2}{\sum(x_i - \bar{x})^2}}$$

- Now we have everything we need to test the hypothesis.
- The normality of the residual terms

- Example: 47 swiss cantons, x=percentage of labor force in agriculture, y=standardized fertility measure.

$$b_0 = 60.30$$

$$b_1 = 0.194$$

$$s = 11.82$$

$$SSX = 23726.77$$

$$s_{b_1} = 11.82/\sqrt{23726.77} = 11.82/154.03 = 0.076$$

$$t = \frac{0.194}{0.076} = 2.55$$

What is the DF for this t-test? (n-2=45).

The exact p-value turns out to be 0.014.

Calculate 95% CI for b_1 .

$$0.194 \pm 2.01 * .076 = (0.041, 0.347)$$

- Analysis of Variance

- We can express our model in a slightly different format:

$$(y_i - \bar{y}) = (\hat{y}_i - \bar{y}) + (y_i - \hat{y}_i)$$

- Put TOTAL, MODEL, and RESIDUAL below. Explain that we are "parsing" the total variation in y.

$$TOTAL = MODEL + RESIDUAL$$

- This is for one individual, but we can also do it for full sample:

$$\sum (y_i - \bar{y})^2 = \sum (\hat{y}_i - \bar{y})^2 + \sum (y_i - \hat{y}_i)^2$$

$$SST = SSM + SSE$$

* $SST = \sum (y_i - \bar{y})^2$ =total sum of squares

* $SSM = \sum (\hat{y}_i - \bar{y})^2$ =model sum of squares

* $SSE = \sum (y_i - \hat{y}_i)^2$ =error sum of squares

- Each portion of the model also has degrees of freedom associated with it.

* $DFT = n - 1$ (sample size minus one for mean)

* $DFM = 1$ (number of non-intercept parameters)

* $DFE = n - 1 - DFM$

$$DFT = DFM + DFE$$

– Dividing SS by DF give you the mean squares, MS.

– Some linkages:

* Our estimate of s^2 is MSE.

$$MSE = \frac{SSE}{DFE} = \frac{\sum(y_i - \hat{y}_i)^2}{n - 2} = s^2$$

* r^2 is given by:

$$\frac{SSM}{SST} = \frac{\sum(\hat{y}_i - \bar{y})^2}{\sum(y_i - \bar{y})^2}$$

– If $H_0 : \beta_1 = 0$ is true, then :

$$\hat{y}_i = \bar{y}$$

so,

$$SSM = 0$$

$$SST = SSE$$

– It turns out we can use this fact to generate another test statistic:

$$F = \frac{MSM}{MSE}$$

– If the null hypothesis is true, F should be zero. Its sampling distribution under the null hypothesis is given by the F-distribution which has two parameters: df_1 and df_2 given by 1 and $n - 2$, respectively.

– Like the other tests, we just ask what is the likelihood of getting a value for F this big or bigger just by random chance.

– Previous Example

Source	DF	SS	MS	F	Pr(> F)
Model	1	894.8	894.8	6.41	0.014
Residual	45	6283.1	139.6		
Total	46	7177.9			

– For a value of 6.41 on an $F(1, 45)$, what is the approximate p-value? In exact terms, it is 0.014, the same as we got for our two-sided test of b_1 in this same example.

– In general, in the bivariate case, the F-test and two-sided t-test will give you the same results.

2.11 Inference for multivariate regression

- Inference for multivariate regression is a straightforward extension of the bivariate case.
- Our population regression equation is given by:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + \beta_3 x_{i3} + \cdots + \beta_p x_{ip} + \epsilon_i$$

Where the variance of the residuals is given by σ^2

- Equation has $p + 2$ total parameters (the $p + 1$ β parameters and the σ term), that must be estimated from sample data:

$$[\beta_0, \beta_1, \dots, \beta_p, \sigma] \rightarrow [b_0, b_1, \dots, b_p, s]$$

- As before, we estimate these β 's by minimizing the squared residual distances.
- σ is still estimated as before:

$$s^2 = \frac{\sum (y_i - \hat{y}_i)^2}{n - p - 1}$$

Except $n - 2$ is now $n - p - 1$ where p is the number of non-intercept β 's.

- We can run a t-test on each individual b_j just as before to test whether it is zero or not.

$$t = \frac{b_j}{s_{b_j}}$$

where the degrees of freedom for the t-test are $n - p - 1$.

- In order to do this, we need to calculate s_{b_j} , which is the estimated standard error for the sampling distribution of b_j . The math here is complicated, so we will not do this calculation in class. Nonetheless, you should have an intuitive understanding of what the value s_{b_j} is.

- The F-test is also similar except that DFM becomes p and DFE becomes $n - p - 1$

Source	DF	SS	MS	F
Model	p	SSM	SSM/DFM	MSM/MSE
Residual	$n - p - 1$	SSE	SSE/DFE	
Total	$n - 1$	SST		

The meaning of the F-test changes however. The H_0 becomes:

$$H_0 : \beta_0 = \beta_1 = \beta_2 = \cdots = \beta_p = 0$$

$$H_a : \text{one of } \beta_j \neq 0$$

- Example: Let's return to our example of abortion attitudes on the GSS.

Variable	Coefficient	SE	t-stat
Intercept	2.82	0.071	39.8
Age (mean centered)	-0.0033	0.0028	-1.17
Education (mean centered)	0.056	0.0198	2.82
Female*Education	0.035	0.028	1.28
Sex			
Male (ref)	-		
Female	0.089	0.083	1.07
Marital Status			
Married (ref)	-		
Widowed	-0.050	0.160	-0.31
Divorced	0.134	0.110	1.21
Never Married	0.394	0.113	3.50
R^2	0.042		

- Which of these parameters can we be certain are not zero at the 5% level?
- Do we have strong evidence that the effect of education differs between men and women?
- What groups can we reasonably say are different from the married group?
- We can also construct an ANOVA table.

Source	DF	SS	MS	F
Model	7	122.1	17.44	8.27
Residual	1322	2787.09	2.11	
Total	1329	1329		

2.12 Non-parametric tests

- The biggest assumption of hypothesis tests is that you know the t statistic is actually distributed as a t -distribution.
- In large samples, you can rely on the CLT.
- Small samples are another issue. If the underlying population distribution of the variable x is normal then you have no worries. In many cases even non-normal population distributions will produce sampling distributions of t that are close to the t -distribution.
- However, in some cases with small samples, you might be better off using **non-parametric tests**.
 - really small samples ($n < 20$)
 - outliers
 - strong non-normality
- Non-parametric tests are not based upon inferring the value of any population parameter.
- For something like a two-sample t -test the most common non-parametric test for continuous data is the **Wilcoxon Rank-Sum Test**.
 - Comparing distributions
 - H_0 : distributions of x for sample A and B are the same.
 - H_a : one distribution has systematically higher values (a location shift)
 - The intuition is to rank the observations from highest to lowest and see if one group is clearly on top.
 - The steps:
 1. rank observations from smallest to largest
 2. produce rank score for each observation (one kind of transformation) - ties should be given average rank
 3. Compare the sum of the ranks between the groups.

If both samples are drawn from the same distributions, then their rank sums should be proportional to their group size.

w = sum of ranks for first group

n_1 = size of first group

n_2 =size of second group

$$N = n_1 + n_2$$

If the two distributions are identical (H_0), then:

$$E(w) = n_1(N + 1)/2$$

$$V(w) = \sqrt{n_1 n_2 (N + 1) / 12}$$

4. The statistic w has a precise sampling distribution. It is a bit unwieldy but can be used for small samples. For big samples the sampling distribution becomes approximately normal, but by then you should be safe with a t-test anyway.

– Example: Life expectancy (2000) in social democratic vs. corporatist nations

Country	type	life exp.	rank
Denmark	(s)	76.5	1
Germany	(c)	77.4	2
Austria	(c)	77.7	3
Belgium	(c)	77.8	4
France	(c)	78.0	5
Netherlands	(s?)	78.3	6
Norway	(s)	78.7	7
Italy	(c)	79	8
Sweden	(s)	79.6	9

$$w(s) = 1 + 6 + 7 + 9 = 23$$

$$E(w) = 4 * 9 / 2 = 18$$

On the table, with $n_a = 4$ and $n_b = 5$, the upper tail requirement for $\alpha = .05$ is a w of 28. Our value of 23 doesn't even reach a p-value of 0.20. So we can't say that the higher rankings of social democratic countries aren't due to random chance.

- There are other examples of non-parametric tests, but this example gives the flavor.

2.13 Catch-up and review