6 Categorical data

So far, we've focused on analyzing numerical data. This section focuses on data that's categorical (e.g., with values like 'red' or 'blue' that don't have any ordering). We'll start with the case where both the inputs and the outputs are categorical, then discuss categorical inputs with numerical outputs.

6.1 Categorical input with categorical output

We can look at a table of counts for each input/output category pair: such a table is usually called a two-way table or a contingency table. Here's an example:

	Outcome 1	Outcome 2
Treatment 1	A	B
Treatment 2	C	D

The letters A, B, C, and D represent numeric counts: for example, A is the number of data points that obtained Treatment 1 and ended up with Outcome 1. Here are some important definitions:

- The risk of Outcome 1 is $A = (A + B)$ for treatment 1 and $C = (C + D)$ for treatment 2: this is the proportion of data points that fall in this category, and can be interpreted as the (empirical) conditional probability of Outcome 1 given Treatment 1.

  For example, suppose the treatments correspond to smoking and non-smoking, and the outcomes correspond to cancer or no cancer. Then, the risk of cancer for smoking is $A = (A + B)$: this matches up with our intuitive understanding of the word "risk".

- The relative risk is $\frac{A/(A+B)}{C/(C+D)}$ Intuitively, this compares the risk of one treatment relative to the other.

- The odds ratio is $\frac{A/B}{C/D}$. Intuitively, this compares the odds of the two outcomes across the two treatments. The odds ratio is useful as a measure of practical significance: while a small effect can be statistically significant, we're often interested in the size of an effect as well as its significance. Using a confidence interval around an odds ratio can help capture this.

6.1.1 Simpson's Paradox

With categorical data, it's extremely important to keep confounding factors in mind! For example, suppose we have data from two hospitals on a risky surgical procedure:

	Lived	Died	Survival rate
Hospital A	80	120	40%
Hospital B	20	80	20%

It seems obvious from this data that Hospital B is worse by a significant margin, and with this many samples, the effect is statistically significant (we'll see how to test this a little later).

Now suppose we learn that Hospital A is in a better part of town than Hospital B, and that the condition of incoming patients might be a significant factor in patient outcome. Here's what happens when we break down the data by patient condition:

	Good condition			Bad condition		
	Lived	Died	Survival rate	Lived	Died	Survival rate
Hospital A	80	100	44%	0	20	0%
Hospital B	10	10	50%	10	70	13%

Suddenly, Hospital B performs better in both cases! This is known as Simpson's Paradox. Intuitively, this happens because of the confounding effect of patient condition: Hospital A appears to do better at first, but all of its survivors are patients in good condition. Once we looked more closely at the data given patient condition, we saw that Hospital A was actually worse, but looked better at first because of having more "good-condition" patients.

EXAMPLE: TITANIC SURVIVAL RATES

The following table shows survival rates from the sinking of the Titanica:

	First class	Second class	Third class	Crew	Total
Lived	203	118	178	212	711%
Died	122	167	528	696	1513%
Survival rate	62%	41%	25%	23%	32%

From this table, it appears that there was a significant class bias in surivors of the Titanic: first-class passengers seemed to survive at a much higher rate. However, once again, the initial table doesn't tell the full story. While the Titanic was sinking, the lifeboats were generally filled with women and children first. As the next table shows, the gender ratios of all passengers between the different classes were dramatically different:

	First class	Second class	Third class	Crew	Total
Children	6	24	79	0	109
Woman	144	93	165	23	425
Men	175	168	462	885	1690

In order to validate this conclusion, we'd need to look at the data broken down by all three factors (gender, class, and surival). In this particular case the data would indeed support that conclusion, but can you come up with an example where the two tables we have here might not be enough?

EXAMPLE: BERKELEY ADMISSIONS, 1973

When UC Berkeley released their graduate school admission numbers in 1973, men and women appeared to be admitted at different rates:

	Outcome 1	Outcome 2
Treatment 1	A	B
Treatment 2	C	D

This led to a lawsuit against the school. However, if we look at the admission rates by department, we can see that some departments were much more selective than others, and tha department selectivity is confounded with gender in the original data:

Department	Men		Women	
	Applicants	Acceptance rate	Applicants	Acceptance rate
A	825	62%	108	82%
B	560	63%	25	68%
C	325	37%	593	34%
D	417	33%	375	35%
E	191	28%	393	24%
F	272	6%	341	7%

After breaking it down by department, we see that most of the men applied to the less selective departments, while most of the women applied to the more selective departments. On top of that, in 4 out of the 6 departments, women were admitted at a higher rate than men! But, when we looked at the original table, we weren't able to see the effect of this confounding variable.

This example illustrates why it's critical to account for confounding factors: a better analysis of the data would look at the data for each department separately, or otherwise include the department in the analysis (we'll see a way of doing this later).

6.1.2 Significance testing: the χ^2 test

So now we’ve acquired our data, made sure we don’t have any bad confounding factors, and we’re ready to analyze it. We’ll focus on the task of measuring whether the inputs had a substantial effect on the outputs.

Our null hypothesis will usually be that the inputs (treatments) don’t affect the output (outcome). We’ll determine the ”expected” count in each entry based on the null hypothesis, and compute the following test statistic:

$$\chi^2 = \sum_{\text{table entries}} \frac{(\text{observed - expected})^2}{\text{expected}} \tag{6.1}$$

If the value is ”big enough”, then we can reject the null hypothesis.

In this equation, the ”observed” counts are simply the data we observe. What does ”expected” mean here? We can’t just divide the total uniformly across all boxes: for example, in the hospital example above, Hospital A saw more patients than Hospital B overall, and more people died than lived overall. Neither of these effects are relevant to the effect of hospital on outcome, so we have to account for them while computing the expected counts. The following table shows the expected counts (rounded to the nearest integer for convenience):

Expected	Lived	Died	Survival rate
Hospital A	2/3·1/3·300=67	2/3·2/3·300=133	33%
Hospital B	1/3·1/3·300=33	1/3·2/3·300=67	33%

In particular, the first row (Hospital A) accounts for 2/3 of the total data, while the first column (survived) accounts for 1/3 of the total data. So, the top left entry (patients from Hospital A who survived) should account for $2/3 \cdot 1/3 = 2/9$ of the total data. Notice that in both hospitals, the survival rate is exactly the same.

In general, suppose we have a table like the one shown below:

		Outcome				
		1	...	j	...	Total
Treatment	1					x_1
	⋮					⋮
	i					x_i
	⋮					⋮
	Total	y_1	...	y_j	...	N

Let x_i be the total for row i, y_j be the total for column j, and N be the total number of entries in the table. Then the expected count for the entry at row i and column j is $\frac{x_i y_i}{N}$:

		Outcome				
		1	...	j	...	Total
Treatment	1					x_1
	⋮					⋮
	i			$\frac{x_i y_i}{N}$		x_i
	⋮					⋮
	Total	y_1	...	y_j	...	N

Alternately, if we view our two-way table as measuring the joint distribution of the input and output variables, then the expected counts are the ones computed as if the two variables were independent.

Going back to the hospital example, the expected counts are 66.7 for Hospital A/lived, 133.3 for Hospital A/died, 33.3 for Hospital B/lived, and 66.7 for Hospital B/died. Using (6.1), we compute $\chi^2 = 12$. How do we determine what this means? As you may have guessed, it follows a $\chi^2 = 12$ distribution under the null hypothesis: we’ll use this fact to look at how likely our observed value is.

We'll assume that each data point was collected independently. Then each table entry is simply a binomial random variable (because it's the sum of a bunch of independent binary variables: "was this subject in this category or not?"). If the numbers in our table are large enough, we can approximate these as Gaussian. Let's take another look at (6.1):

$$\chi^2 = \sum_{\text{table entries}} \frac{(\text{observed - expected})^2}{\text{expected}}$$

So, for each table entry, we're taking a Gaussian random variable and just standardizing it (i.e., subtracting the mean and dividing by the standard deviation). Since we're squaring them and adding them up, the result is defined to be a χ^2 random variable. So, we can just run a χ^2 test; if there are r rows and c columns in the table, then the test statistic distribution has $(r-1)\cdot(c-1)$ degrees of freedom. In the hospital example, we would have obtained a p-value of 0.00053 for our initial table (because this p-value is so small, this suggests that we should reject the null hypothesis that which hospital a patient goes to has no effect on the patient's outcome).

Notice the two assumptions we made: independent data points and large enough values at each entry. As a rule of thumb, at least about 10 samples are usually needed in each entry for the approximation to work properly.

What if the numbers in our table aren't big enough? If the table is 2×2, then we can use Fisher's Exact Test. This is essentially a permutation test (as in Chapter 5), and it computes the exact probability of obtaining our particular arrangement of the data under the null hypothesis that the outputs and inputs are independent. We can directly compute the p-value (known as the Fisher p-value). Assuming entries A, B, C, and D as earlier, then we get:

$$p = \frac{\binom{A+B}{A}\binom{C+D}{C}}{\binom{N}{A+C}},$$

where $\binom{n}{k} = \frac{n!}{k!(n-k)!}$. Notice that if the table entries are small, computing these factorials isn't too troublesome. When the numbers get bigger, it's often more difficult to compute numerically, but in such cases the Gaussian/χ^2 approximation usually works well.

If the table is larger, we can use the Yates correction, which simply subtracts 0.5 from all our counts.(Why?) This makes the Gaussian approximation slightly more accurate. We can also run something similar to Fisher's test which simulates the Fisher p-value by randomly sampling different permutations of the table instead of computing it exactly as above. Your software will usually have an option you can set for this (usually called a simulated p-value or Monte Carlo approximation to Fisher's test).

We can also use these tests with other null hypotheses: if we want to compare two sets of points, or compare to a null hypothesis other than output/input independence, we can simply adjust the expected counts accordingly and do everything else just as before.

6.2 Categorical inputs with continuous outputs: ANOVA

So far we've talked about the case where both our input and output are categorical. What if our outputs are continuous? ANOVA (which stands for ANalysis Of VAriance) is a technique for testing whether the continuous outputs depend on the inputs. Equivalently, it tests whether different input categories have significantly different values for the output variable.

A quick aside about vocab: categorical variables are called factors and the values they take on are called levels. For example, a factor might be "color" and its levels might be "red", "blue", and "yellow". These different possibilities are also often called groups: that is, we might talk about the "red" group (i.e., all the red data points), the “blue” group, and so on.

We'll present two complementary perspectives on ANOVA to help provide some intuition on what it's doing as well as how it's doing it. The first section sets up the model for ANOVA and describes the test, while the second describes how ANOVA can be viewed as a form of linear regression.

6.2.1 ANOVA as comparing means

Suppose our input factor has k different levels (remember, this means k different possible values it can take on). For simplicity, we'll assume the categories are named 1,2,...,k. We'll call the input factors for each data point $x_1, ..., x_n$ for our data. We'll also assume we have some continuous output variable which we'll call $y_1, ..., y_n$. As an example, suppose our input variable is one of five medical treatments and our output variable is weight in pounds. Then x_1 might be 4 (corresponding the fourth treatment) and y_1 might be 150.

The model for ANOVA assumes that there are k group-specific means; we'll call these $\mu_1, ...\mu_k$. The group mean for point i is then μ_x (since x_i is a number between 1 and k). This suggests that we can write each output data point y_i as $y_i = \mu_{xi} + \epsilon_i$, where ϵ_i is random noise. We'd then test whether all the means μ_k are equal (more on how exactly to do that in the next section).

To be a bit more exact, we'll usually break down μ_{x_i} and write it as $\mu_{xi} = \mu + \tau_{x_i}$, where μ is a global mean and τ is a group-specific offset. Our model is then

$$y_i = \overbrace{\underbrace{\mu}_{\text{global mean}} + \underbrace{\tau_{x_i}}_{\text{group-specific offset}}}^{\text{mean for this point's group}(=\mu_{x_i})} + \underbrace{\epsilon_i}_{\text{random noise}} \tag{6.2}$$

The test then reduces to seeing whether the offsets τ are all 0. The null hypothesis in one-way ANOVA is that the group means $\mu_1, \mu_2, ..., \mu_k$ are all equal.

6.2.2 ANOVA as linear regression

Suppose again that our input factor has k different levels. We'll represent this numerically with a binary vector with exactly one 1. What does that mean? For example, suppose again that our input can be "red", "blue", or "yellow". Then we'll represent "red" with (1 0 0) "blue" with (0 1 0), and "yellow" with (0 0 1). For example, if we have 5 data points with input values red, red, blue, yellow, and blue, respectively, we'd use the following matrix to represent the data:

$$X = \begin{pmatrix} 1 & 0 & 0 \\ 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 0 & 1 \\ 0 & 1 & 0 \end{pmatrix}$$

This provides us with a perfect input for multiple regression. Once we run regression, we can see if the model is a good fit: if it is, then the categorization from the input X is a good way to explain the output. This means that there is some significant difference between the groups. When discussing linear regression, we learned that the best way to evaluate how well the model fits is to use the F-test, comparing "the variance explained by the model" to "the variance not explained by the model". This is exactly how one-way ANOVA works! Since we already understand linear regression, we can view it from this perspective and everything we've already learned about regression carries over. Note that ANOVA is often expressed in terms of comparing variance within groups to variance between groups, which is an equivalent way of doing the same thing.

The output of running ANOVA in your software will be an F-statistic and a p-value. The F-statistic is computed the same way we did it in Chapter 3: we can decompose the total variability in the data into $TSS = ESS + RSS$, giving us $F = ESS/RSS$ with proper normalizers with respect to degrees of freedom.

6.2.3 ANOVA: Interpretations and assumptions

Note that this kind of test only tells you that there's some difference between the groups; it doesn't necessarily tell you where that difference comes from! In order to find which ones are different, we'll have to do post hoc tests, such as t-tests between pairs of groups. When doing tests like this, it's important to remember the multiple comparison corrections we looked at in Chapter 2. While we could have just done all the pairwise tests to begin with, ANOVA helps us prevent false positives.

What assumptions does ANOVA make? Since we know ANOVA is really just a specific case of linear regression, we can list out the assumptions of linear regression and see that they carry over:

- Identical variance between groups: the variance from the group mean for each group is the regression residual. We know from before that the residual variance for all points is the same! Notice that ANOVA is particularly sensitive to this assumption: if the data are heteroscedastic (i.e., the groups have different variances), ANOVA-based tests will often fail.
- Points are normally distributed and independent: the outputs in regression are assumed to have independent Gaussian distributions centered around the prediction $X\beta$, so the same must be true of the output values for ANOVA.

6.2.4 Some extensions of ANOVA

Two-way ANOVA

If we're interested in measuring the effects of two input factors, we'll use a two-way ANOVA. If we call the second input factor $z_1, ..., z_n$, then our model from Equation (6.2) becomes:

$$y_i = \overbrace{\underbrace{\mu}_{\text{global mean}} + \underbrace{\tau_{x_i}}_{\text{offset for factor 1+}} + \underbrace{\eta_{x_i}}_{\text{offset for factor 2}} + \underbrace{\gamma_{x_i z_i}}_{\text{offset for interactions}}}^{\text{mean for this point's group}} + \underbrace{\epsilon_i}_{\text{random noise}} \tag{6.3}$$

This looks a little more complex: we've added the term η_{z_i} to account for the second factor, but we also added an interaction term. This is useful for modeling effects that wouldn't happen with either of the two individual values alone, but happen due to interactions between them. For example, in an experiment involving multiple treatments for cancer, it's possible that while each of two treatments work well individually, the combination of the two cancels them out. In this case, the corresponding term would be negative.

In our regression formulation, we'll now have one set of predictors in X for each input, and we can also add predictors corresponding to interactions between the two.

For example, suppose the first input variable is color as before, and the second is shape: "circle", "square", "triangle", or "star". If our data points are "red square", "red star", "yellow triangle", and "blue star", then we'd use the following inputs :

$$X = \left(\begin{array}{ccc|cccc} & \text{Color} & & & \text{Shape} & & \\ \text{R} & \text{B} & \text{Y} & \bullet & \blacksquare & \blacktriangle & \bigstar \\ 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right)$$

Here, the vertical line just marks the boundary between color features and shape features: we'd still use the entire matrix when computing the regression coefficients.

Warning: in a two-way ANOVA, it's important to make sure the counts for all category pairs are reasonably large! The example above violates this particularly egregiously. For example, suppose we analyze this data and obtain a large coefficient (i.e., the coefficient for "yellow" is particularly large). However, we can't tell whether this effect is because of the color yellow or the shape triangle (or worse still, because of the interaction between the two), since our only yellow data point is also our only triangle.

So, what does the two-way ANOVA give us? Suppose we have two input variables, A and B. We'll further decompose ESS as $SS_A + SS_B + SS_{AB}$, where SS_A and SS_B come from only taking the predictors/coefficients corresponding to input variables A and B respectively. Similarly, SS_{AB} comes from only taking the interaction predictors/coefficients. In the example above, let X_A be the first 3 columns of X, β_A be the first 3 elements of β, and compute SS_A from $X_A\beta_A$. We can then compute F statistics for each of SS_A, SS_B, and SS_{AB}, which tell us about the effect of A and B individually, and the interaction effect respectively.

This method can be generalized to multiple factors: it's not uncommon to see 3-way ANOVAs and higher order analyses.

ANCOVA: Analysis of covariance

Suppose we have both continuous and categorical inputs: this might happen in a case where we're interested in the effect of both continuous and categorical inputs on our output. Alternately, we may want to control for (continuous) external sources of variation that we aren't interested in! For example, we may want to know the efficiency of a drug, but want to control for the effect of patient age on outcome. In order to control for this effect, we'll use ANCOVA.

By viewing ANOVA as linear regression, adding continuous factors is easy: all we have to do is add another column to our input matrix X! We'll compute ESS as $SS_{interest} + SS_{nuisance}$, where "interest" and "nuisance" correspond to the variables we want to measure and the ones we're controlling for, respectively. We'll break things up the same way we did in two-way ANOVA. When performing our F-test, we'll only look at $SS_{interest}$, since we're not interested in the variability from the nuisance components.

MANOVA: Multiple ANOVA

Multiple ANOVA is used when we have multiple outputs which may or may not be independent. This can be viewed as an extension of General Linear Models (which are an extension of linear regression to the multi-output case). Similarly, MANCOVA is used when we need to control for nuisance factors and we have multiple outputs.

6.2.5 ANOVA and statistical software

Notice that although we've learned that we can think of and understand ANOVA as simply a type of linear regression, it's often a good idea to use ANOVA-specific options in your statistical software of choice: they'll often have better options for specifying things like ANCOVA covariates and will present the output in a more appropriate way.

6.2.6 Kruskal-Wallis: the non-parametric version

In Chapter 5, we saw the Wilcoxon signed-rank test and the Mann-Whitney U test. These were useful for comparing medians of two groups when the t-test's assumptions about normally distributed means didn't apply. Similarly for ANOVA, the Kruskal-Wallis one-way analysis of variance is designed to compare the medians of several groups. In particular, the null hypothesis of the Kruskal-Wallis test is that the medians of several groups are the same. While this test doesn't assume Gaussian distributions within each group, it does assume that the distribution for each group has the same shape. For example, if all the groups being compared are heavily skewed in the same direction, Kruskal-Wallis would be an appropriate alternative to ANOVA. However, if they skew in different directions or have greatly different distributions, then this test may not be appropriate.

6.3 Summary: statistical tests

We've seen a lot of statistical tests so far, from the t-test to the F-test to the ANOVA family. Each of these tests is useful in the situation it was designed for, and often inappropriate in other settings. When choosing the right test for your data, it's important to keep things like variable type (numeric, categorical, or ordinal), sample sizes, and distributional assumptions in mind. Most of the tests described here have particular assumptions, and while these assumptions are often quite general (such as in the case of the t-test for a reasonably sized sample), they are required for the output of the test to be valid and interpretable.

7 Experimental Design

This chapter focuses on how to design experiments. We bring the assumptions we've learned into play and discuss key ideas and principles of good experimental design. Good design concepts are best illustrated by examples, so we provide many in this chapter, mostly involving clinical trials or improving education in developing countries.

7.1 Core elements of good design

While experimental design is a broad topic that is often difficult to get right, there are a few guiding principles that all good designs are built on top of:

1. Replication: Any good experiment should be reproducible, and in particular, replication should yield similar results. Shockingly, many published scientific papers fail at this tenet! (This is one of the greatest secrets of academia.) Meanwhile, anecdotal evidence is not scientific proof, and we've seen so far that most of the methods we've discussed improve as the number of samples increases. (This applies only when the initial model is in a ballpark of the true model. Otherwise, simplying adding data won't change the model performance.) This is often difficult to achieve because of cost or time constraints in a study: while gathering an infinite amount of data might be theoretically ideal, it's practically impossible.

2. Comparison/control/baseline: In any experiment where you're measuring the effect of a treatment, it's impossible to assess that effect without having a reference value.

   For example, a remedial summer program in a poor school district might result in little to no improvement in student performance between the beginning and end of the summer. But what would have happened if we hadn't intervened? In disadvantaged schools, academic performance can sometimes decline over the summer - in this case, our intervention would be classified as an improvement!

   As another example, when an experimental medical treatment undergoes clinical trials, it is standard to compare the treatment against a placebo, which refers to a sham treatment that is advertised as effective. In these trials, the goal is to show that the experimental treatment significantly outperforms the placebo. The exception to using a placebo in these cases is when the disease is extremely debilitating where it makes sense to compare against a standard treatment currently in use (but that perhaps isn't very effective) rather than a placebo. Despite a placebo being a sham treatment, it can often actually make a subject feel better compared to not giving any treatment at all! This is explained in the next example panel.

EXAMPLE: THE PLACEBO AND HAWTHORNE EFFECTS

Beware the placebo effect! When you apply a treatment to a subject, that treatment may be ineffective, but can still produce a significant effect simply due to the existence of the treatment. For example, a fake placebo surgery can actually do as well as a common knee surgery! And giving people nonalcoholic drinks but telling them that the drinks are alcoholic can result in a decline in memory powers!

Because of the placebo effect, development of medical treatments demands the stronger standard of outperforming a placebo rather than outperforming not giving any treatment at all since a placebo alone could already result in a startling improvement in a test subject's condition, often largely due to psychological factors.

To examine the benefit of a placebo, an experiment could have a control group that receives no treatment, a placebo group that receives a sham treatment, and a treatment group that recieves the actual treatment under study. By doing this, the experimenter can measure both the effect of the placebo over the baseline and the effect of the treatment over the placebo.

Closely related to the placebo effect is the Hawthorne effect: in a behavioral study, the behavior of subjects might be due to their reaction to being studied. In a famous experiment in the early 20th century, a factory called the Hawthorne Works wanted to measure the effect of lighting on productivity. An experimental group had their light bulbs changed, and the experimenters wanted to measure the effect on productivity. A control group saw workers change their bulbs, but the new bulbs were identical to the old ones. However, both groups improved after the "new" bulbs were put in: the control group improved purely due to their perception of an effect.

These effects can turn up where you least expect them to! For example, suppose you're designing an experiment to measure the effect of fertilizers on a farm. While the plants probably aren't vulnerable to the placebo effect, the farmers could be. A farmer whose field is fertilized might work harder and be more motivated simply by being part of the study. As a result, in such a study, it might be a good idea to have a placebo farm that receives plain dirt. There are also other confounding factors: seemingly uninteresting quantities such as the overall moisture level of the fertilizer may have an impact on the result, so it's important to make the placebo group as equalized as possible!

3. Blocking/controlling for confounds: In an experiment with possible external sources of variability, it's always best to control for these factors (recall Simpson's paradox from Chapter 5!). Controlling for confounds is best achieved through blocked design, where we divide subjects into groups corresponding to levels of a confounding factor, and repeat the experiment for each group. By accounting for the effect of confounding variables, we can avoid being misled by our data.

   For example, an educational intervention program may have different effects on students of different gender. If we aren't interested in the confounding effect of gender, then we can analyze the blocks separately or even include this confound in our analysis (e.g., using ANCOVA).

4. Randomization: Most theoretical analysis assumes that data points are independent. Randomization is often the key ingredient to satisfying this assumption! For example, it's better to have randomly selected data points, and to randomly assign those data points to different groups/treatments, and so on. Skipping randomization can often lead to bias in data!

A general rule of thumb, attributed to the famous statistician George Box, is to "block what you can, randomize what you cannot."

We'll see concrete examples of how these principles come into play.

7.2 Gathering Samples

A critical part of any experiment is gathering samples or data points. In all of these examples, we assume some underlying "population". For example, if you're conducting a poll for all of the U.S., then your population could be all U.S. residents. If you're studying the effect of a new pilot bilingual immersion program at a high school, then your population would be students from that high school. Here are a few ways to gather data:

1. Simple Random Sample (SRS): In a simple random sample, we draw members of the population uniformly at random without replacement. This is like putting the names of everyone in the population into a hat and then drawing a few names out of the hat, assuming of course that the drawing is fair. The "without replacement" part just means that once a name has been drawn from the hat, we don't put it back into the hat. Effectively we can't draw the same name twice. We'll see in Section 7.3 that an SRS consists of points that are not independent! However, if the number of data points in the SRS is much smaller than the total population size, then we can safely treat the samples as approximately independent.

   Unfortunately, collecting an SRS is often difficult to carry out in practice: if we want to randomly sample the population of people in New York City with landline phones, then we can take a phonebook, choose names at random, and call them. But, if we're sampling the population of students in a developing country, it's almost impossible to find a list, let alone obtain access to people that are uniformly sampled.

   Another issue with simple random samples is that it's often difficult to make conclusions about smaller subpopulations. For example, if a particular subgroup is relatively small, a uniform sample may not capture any members of that group. For example, if we're trying to estimate the proportion of the population that has an extremely rare disease (say, 1 in a million), then chances are that from sampling, say, 1000 people, none of them are going to have the disease.

2. Stratified Random Sample: (First off, by convention, the abbreviation "SRS" refers to a simple random sample, and not a stratified random sample!) As for stratified random samples: suppose we know that the population consists of several different non-overlapping groups, and that there isn't much variation within each group. Then we can divide the population into these groups and within each group collect an SRS. These groups are called strata, with each group called a stratum.

   In a scheme known as proportional allocation, the number of subjects per stratum is usually chosen to match that stratum's true relative size in the population. For example, if 60% of the population of interest is female, and we have two strata, one per gender, then if we want our study to have a total of 1,000 people, proportional allocation would ask that we collect two separate SRSs, one with 600 women and one with 400 men.

   In Neyman's optimal allocation, the number of subjects per stratum is determined by both the stratum's true relative size in the population as well as the variance within the stratum. If either the relative size of the stratum in the population is larger, or the variance within the stratum is higher, then we'll collect more samples from this stratum. Formally, if W_ℓ is the true proportion of the population that is in stratum ℓ, and σ_ℓ is the true standard deviation within stratum ℓ, then Neyman's optimal allocation says that the size of the SRS for stratum ℓ should be proportional to $W_\ell \sigma_\ell$.

   This technique allows us to accurately measure the effects of small groups that may have otherwise been missed in an SRS over the whole population (i.e., without stratification). For example, we may want to sample the performance of students in different types of schools. If some school categories are larger (i.e., have more students) than others, then an SRS over the whole population may miss the small categories. A stratified sample would list the categories and sample randomly within each type of school.

3. Cluster Sampling: The two methods above require samples from either the entire population or every single stratum. This may not always be cost-effective or even feasible. Cluster sampling is based on the idea of dividing the population into natural, heterogeneous groups that are relatively similar to each other. Each group should be well-representative of the population. Instead of sampling from all of the groups, we'll randomly sample a few, and then do random sampling within each one. Since they're all similar to each other, then a random sample from one should be representative of a random sample of the population.

   For example, if we're polling a city, we might divide it up into city blocks. Then we randomly choose some number of blocks to sample. Finally, within each block, we collect an SRS. As long as there are no large differences between each block, and each block represents the overall city population well, then this technique is often more cost-effective than an SRS over the whole population.

However, all of these frameworks have issues:

- Getting an unbiased list of subjects to sample from, even within a stratum or a cluster, can often be difficult.
- We may have non-response bias: in a study of people such as a survey, there will almost always be people who choose not to respond. Unfortunately, different groups often have different non-response rates. For example, in an approval survey, more enthusiastic people are more likely to respond to questions, which can bias the results toward the extremes. As another example, a poll asking about the workload of students may run into non-response bias where overworked students are too busy to respond and, as a result, the collected responses may suggest that people work fewer hours than they actually do on average.
- For surveys, how questions are worded can make a huge difference in how people respond! We see this in the following example.

EXAMPLE: WORDING MATTERS!

In October 2004, Stanley Presser ran a poll for The New Yorker, where half of respondents were asked "Do you think the United States should allow public speeches against democracy?" and the other half were asked the same question except with "allow" replaced by "forbid". Whereas 56% answered no to "forbidding", 39% answered yes to "allowing" despite the two answers corresponding to the same response.

Generally speaking, it is a good idea to word questions as neutrally as possible, and if the questions don't have some order dependence, to randomize their ordering.

The above issues often make it hard to extend conclusions beyond a study: any analysis we can do is only valid for the population that we sampled from. This is highlighted by the following example, which led to the downfall of the magazine The Literary Digest.

EXAMPLE: THE LITERARY DIGEST POLL FOR 1936 US PRESIDENCY

Republican candidate Alfred Landon was running against Democrat Franklin Delano Roosevelt. The Literary Digest projected that Landon would win by a huge margin: a 57% to 43% victory. The magazine had polled 10 million people and received a whopping 2.4 million responses! Yet Franklin Roosevelt won by a landslide, carrying 46 states while Landon only carried 2 states. The win wasn't just in the electoral college either: Roosevelt won 61% of the popular vote.

What had happened? Of course, a mix of things happened including non-response bias and likely wording issues, but the main issue was selection bias: the questionnaires were sent out to readers of The Literary Digest, those that were in a phone listing, and those on a listing of car owners. But all these lists contain more rich people than poor people, which led to a heavily skewed poll result.

In contrast, a Gallup poll that same year predicted that Roosevelt would receive 56% of the popular vote using only a sample size of 50,000, which turned out to be far more accurate than The Literary Digest's poll results.

7.3 Simple Random Samples: are samples really independent?

Suppose we have a population of N people and we're measuring their heights. Let the true population mean height be μ with variance σ^2. We draw a simple random sample (SRS) of size n from the population. Let $x_1, x_2, ..., x_n$ be the heights we measure of n people in the SRS. Let the sample average height be

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

First off, note that the x_i's are not independent! The intuition is as follows. Let's say the first person we sample from the population is Alice, so x_1 is Alice's height. Since we are sampling without replacement, the second height x_2 cannot come from Alice! So how we selected x_1 affects how we select x_2, which means they're not completely independent. This reasoning extends to all the x_i's.

The sample average (which, as a reminder, is a random variable) turns out to have mean and variance given by

$$\begin{aligned} \mathbb{E}[\bar{x}] &= \mu, \\ Var[\bar{x}] &= \frac{\sigma^2}{n} \underbrace{\left(\frac{N-n}{N-1}\right)}_{\text{correction factor}}, \end{aligned}$$

where we note that there is a correction factor due to the samples not being independent.

If $n \ll N$, then the correction factor is approximately 1, and furthermore, the samples can be approximated as independent and as if they were drawn with replacement. The intuition is that if we're drawing n names from a hat of N names, and $n \ll N$, then even if each time we draw a name, we put the name back into the hat, the chance of us drawing the same name twice is negligible - whether we put the name back in the hat or not doesn't really affect the result!

Confidence intervals

For sufficiently large n, the sample average $\bar{x}$ is approximately normal, but not because of the central limit theorem which we saw earlier (recall that the theorem required that the random variables we're summing to be independent, which isn't the case here). Instead, one needs a fancier central limit theorem that tolerates some dependence between random variables. In any case, the approximate normality of $\bar{x}$ allows us to construct an approximate confidence interval for μ.

As with our earlier excursions into computing confidence intervals, what we need is an estimate of the standard error of our estimator $\bar{x}$. It turns out that an unbiased estimator for $Var[\bar{x}]$ is:

$$\frac{s^2}{n}\left(\frac{N-n}{N}\right)$$

Thus, an estimate of the standard error is

$$s_\mu = \frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}},$$

from which we derive a 95% confidence interval for the mean height μ:

$$\bar{x} \pm 2s_\mu = \bar{x} \pm 2\frac{s}{\sqrt{n}}\sqrt{\frac{N-n}{N}},$$

As a reminder, the coefficient 2 comes from the fact that within 2 standard deviations of a standard normal random variable lies 95% of the probability mass centered around the mean.

If instead the x_i's had been binary random variables taken on values 0 or 1 (e.g., we ask each of n people a yes/no question, where "yes" is encoded as a 1), then one could show that an approximate 95% confidence interval is

$$\bar{x} \pm 2\sqrt{\frac{\bar{x}(1-\bar{x})}{n-1}\cdot\frac{N-n}{N}},$$

where in this context, note that $\bar{x}$ estimates the fraction of the population that has value 1 (e.g., the proportion of people who answer "yes" to a poll).

7.4 Some sample designs

This section covers more experimental designs that are useful for more complex experiments.

7.4.1 Paired tests and repeated measures

Whenever possible, if applying a treatment, it's best to have paired data, where we obtain measurements for each subject before and after treatment. As we saw with t-tests, paired tests often give us the most power.

A generalization of paired tests is repeated measures design. In such a design, we may have multiple (i.e., 2 or more) treatments, and each subject will receive all the treatments. This way, each subject can be thought of as its own control.

For example, suppose we measure the effect of caffeine (in the form of tea and coffee) on student performance. In a repeated measures design, each student would spend a month drinking coffee, a month drinking tea, and a month with no caffeine intake (for control). We may also want to add a month with a decaffeinated drink as a placebo. In such designs, it's important to randomize the order, and to be wary of temporal effects. In this example, stopping caffeine treatment might lead to worse performance due to withdrawal. As a result, it might be worthwhile to wait in between each "measure". We can sometimes model these temporal effects with autocorrelation models, where the errors are no longer assumed to be independent, but rather to depend on each other in sequence.

7.4.2 Randomized complete block design

What do we do when we have multiple factors to block on? If the factors don't depend on each other, then we'll probably have the same number of sub-blocks with in each block. For example, in an experiment where we block on gender and handedness (left or right), we'll have left-handed and right-handed groups for men, and left-handed and right-handed groups for women. Such a design is called complete, because each sub-block is being tested. We'll focus here on cases where we have two blocking factors, although the ideas we'll discuss can be generalized. In a randomized complete block design, we may not have enough data points to replicate within sub-blocks, so we must assign different sub-blocks to different treatment conditions.

Example

For example, suppose we want to measure the effect of giving tablets to students in developing countries. Our experimental condition might be providing students with tablets and giving them an extra hour every day to use them. We would need a control group that receives the normal curriculum, and a placebo group that receives an extra hour of unstructured time (but no tablets) every day. This gives us three levels for the treatment factor: tablet (T), unstructured hour (U), and control (C). Suppose this is a one-year study where we have three terms (fall, spring, and summer), and three (mostly-similar) schools in which to run the experiment. Such a setup is known as a row-column design, and the experimental setup can be illustrated by the following. First, we'll fill in each entry with the treatment we use for that setup. A first attempt at this design (where T, U, and C stand for tablet, unstructured hour, and control respectively) might look like this:

		Time of year		
		Fall	Spring	Summer
Location	1	T	U	C
	2	T	U	C
	3	T	U	C

Unfortunately, this design does not properly take into account the time of year: if we were to run the experiment and see a significant improvement from the tablets, it might have been entirely due to the confounding effect of having the tablet conditions all in the fall! As a result, our ideal design would have each condition appear exactly once per row and once per column (like a Sudoku). Grids that satisfy constraints like this are called latin squares, and we can produce one by taking the table above and shifting each row:

		Time of year		
		Fall	Spring	Summer
Location	1	T	U	C
	2	C	T	U
	3	U	C	T

This way, we'll try each experimental condition in every location and during every time of year.

EXAMPLE: HOW HARD IS EXPERIMENTAL DESIGN, REALLY?

Let's take a seemingly simple example, and see how complicated things can get. Suppose we want to bake the best possible loaf of bread. After some preliminary experimentation, we come up with 2 brands of our, 2 brands of yeast, and 3 oven temperatures, and want to find the optimal combination (out of the 12 possibilities). We find 5 volunteer chefs willing to bake the bread, and 20 volunteer tasters willing to help us evaluate how good it tastes.

Exercise: How would you design an experiment to find the best combination of conditions?

Exercise: Suppose you have each chef bake each of the 12 loaves 4 times (to do this, you'd probably have to upgrade them from volunteeres to paid experimenters!). What are sources of variability within

(a) one loaf of bread?

(b) two loaves with the same recipe and ingredients from the same chef?

(c) two loaves with the same recipe and ingredients from different chefs?

How might you account for these sources of variability?