Chapter 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**[1]. 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 Titanic[a]:

	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:

[1] Note that this isn't really a paradox: while it initially seems impossible, it happens because of a confounding factor.

	First class	Second class	Third class	Crew	Total
Children	6	24	79	0	109
Women	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?

[a]Data source: http://www.anesi.com/titanic.htm

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[a]:

	Total applicants	Acceptance rate
Men	8442	44%
Women	4321	35%

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 that 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).

[a]See Bickel, P.J., et al. Sex Bias in Graduate Admissions: Data from Berkeley. Science, 1975.

■ 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:

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

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	$2/3 \cdot 1/3 \cdot 300 = \mathbf{67}$	$2/3 \cdot 2/3 \cdot 300 = \mathbf{133}$	33%
Hospital B	$1/3 \cdot 1/3 \cdot 300 = \mathbf{33}$	$1/3 \cdot 2/3 \cdot 300 = \mathbf{67}$	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	$\cdots$	j	$\cdots$	Total
Treatment	1					x_1
	$\vdots$					$\vdots$
	i					x_i
	$\vdots$					$\vdots$
	Total	y_1	$\cdots$	y_j	$\cdots$	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 $\frac{x_i y_j}{N}$:

		Outcome				
		1	$\cdots$	j	$\cdots$	Total
Treatment	1					x_1
	$\vdots$					$\vdots$
	i			$\frac{x_i y_j}{N}$		x_i
	$\vdots$					$\vdots$
	Total	y_1	$\cdots$	y_j	$\cdots$	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 χ^2 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):

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

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[2]. 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.

[2]Note that since the entries technically aren't independent, this doesn't exactly meet the requirements of the χ^2 random variable definition. But usually when the numbers in our table are large enough, they can be well-approximated independently.

If the table is larger, we can use the **Yates correction**, which simply subtracts 0.5 from all our counts. 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 $1, 2, \ldots, k$. We'll call the input factors for each data point $x_1, \ldots, x_n$ for our data. We'll also assume we have some continuous output variable which we'll call $y_1, \ldots, y_n$. 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 $\mu_1, \ldots, \mu_k$. The group mean for point i is then μ_{x_i} (since x_i is a number between 1 and k). This suggests that we can write each output data point y_i as $y_i = \mu_{x_i} + \epsilon_i$, 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 $\mu_{x_i} = \mu + \tau_{x_i}$, where μ

is a global mean and $\tau_\cdot$ 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 $\mu_1, \mu_2, \ldots, \mu_k$ 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 $\begin{pmatrix}1 & 0 & 0\end{pmatrix}$, "blue" with $\begin{pmatrix}0 & 1 & 0\end{pmatrix}$, and "yellow" with $\begin{pmatrix}0 & 0 & 1\end{pmatrix}$. 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[3]:

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

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 $\text{SS}_{\text{total}} = \text{SS}_{\text{model}} + \text{SS}_{\text{error}}$, giving us $F = \text{SS}_{\text{model}}/\text{SS}_{\text{error}}$.

■ 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

[3]We also need to include a constant predictor/column to model the the fact that our data usually has some offset or intercept (which we took care of with the global mean μ in Equation (6.2)), but we'll keep things simple and ignore that in these examples.

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 $z_1, \ldots, z_n$, then our model from Equation (6.2) becomes:

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

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 γ 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 :

$$X = \begin{array}{c} \begin{matrix} \multicolumn{3}{c}{\text{Color}} & \multicolumn{4}{c}{\text{Shape}} \\ R & B & Y & \bullet & \blacksquare & \blacktriangle & \bigstar \end{matrix} \\ \left(\begin{array}{ccc|cccc} 1 & 0 & 0 & 0 & 1 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 & 0 & 1 \\ 0 & 0 & 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 & 0 & 1 \end{array}\right) \end{array}$$

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 β_3 (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 SS_{model} as $\text{SS}_A + \text{SS}_B + \text{SS}_{AB}$, 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 SS_{model} as $\text{SS}_{\text{interest}} + \text{SS}_{\text{nuisance}}$, 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 $\text{SS}_{\text{interest}}$, 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 nonparametric 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.