Statistics Consulting Cheat Sheet

Kris Sankaran

October 1, 2017

Contents

1 What this guide is for **3**

2 Hypothesis testing **3**
2.1 (One-sample, Two-sample, and Paired) t-tests 4
2.2 Difference in proportions 6
2.3 Contingency tables 6
2.3.1 χ^2 tests 7
2.3.2 Fisher's Exact test 8
2.3.3 Cochran-Mantel-Haenzel test 8
2.3.4 McNemar's test 9
2.4 Nonparametric tests 9
2.4.1 Rank-based tests 9
2.4.2 Permutation tests 11
2.4.3 Bootstrap tests 12
2.4.4 Kolmogorov-Smirnov 13
2.5 Power analysis 14
2.5.1 Analytical 15
2.5.2 Computational 15

3 Elementary estimation **16**
3.1 Classical confidence intervals 16
3.2 Bootstrap confidence intervals 16

4 (Generalized) Linear Models **17**
4.1 Linear regression 18
4.2 Diagnostics 21
4.3 Logistic regression 22
4.4 Poisson regression 25
4.5 Psueo-Poisson and Negative Binomial regression 27
4.6 Loglinear models 28
4.7 Multinomial regression 29
4.8 Ordinal regression 30

5 Inference in linear models (and other more complex settings 32
5.1 (Generalized) Linear Models and ANOVA . . . 32
5.2 Multiple testing . . . 35
5.2.1 Alternative error metrics . . . 35
5.2.2 Procedures . . . 35
5.3 Causality . . . 36
5.3.1 Propensity score matching . . . 36

6 Regression variants 36
6.1 Random effects and hierarchical models . . . 36
6.2 Curve-fitting . . . 36
6.2.1 Kernel-based . . . 37
6.2.2 Splines . . . 37
6.3 Regularization . . . 37
6.3.1 Ridge, Lasso, and Elastic Net . . . 37
6.3.2 Structured regularization . . . 37
6.4 Time series models . . . 37
6.4.1 ARMA models . . . 37
6.4.2 Hidden Markov Models . . . 37
6.4.3 State-space models . . . 37
6.5 Spatiotemporal models . . . 37
6.6 Survival analysis . . . 37
6.6.1 Kaplan-Meier test . . . 37

7 Model selection 37
7.1 AIC / BIC . . . 37
7.2 Stepwise selection . . . 37
7.3 Lasso . . . 37

8 Unsupervised methods 38
8.1 Clustering . . . 38
8.2 Low-dimensional representations . . . 41
8.2.1 Principle Components Analysis . . . 41
8.2.2 Factor analysis . . . 41
8.2.3 Distance based methods . . . 41
8.3 Networks . . . 41
8.4 Mixture modeling . . . 41
8.4.1 EM . . . 41

9 Data preparation 41
9.1 Missing data . . . 41
9.2 Transformations . . . 41
9.3 Reshaping . . . 42

10 Prediction **42**
10.1 Feature extraction 42
10.2 Nearest-neighbors 42
10.3 Tree-based methods 42
10.4 Kernel-based methods 42
10.5 Metrics 42
10.6 Bias-variance tradeoff 42
10.7 Cross-validation 42

11 Visualization **42**

12 Computation **43**
12.1 Importance sampling 43
12.2 MCMC 43
12.3 Debugging peoples' code 43

1 What this guide is for

- It's hard (probably impossible) to be familiar with all problem types, methods, or domain areas that you might encounter during statistics consulting...so don't try to learn all problem types, methods or domains.
- Instead, try to build up a foundation of core data analysis principles, which you can then adapt to solving a wide variety of problems.
- This doc gives a brief introduction to some of the principles I've found useful during consulting, and while it's no substitute for actual statistics courses / textbooks, it should at least help you identify the statistical abstractions that could be useful for solving client problems
- Finally, there is a lot to consulting outside of pure statistical knowledge – see our tips doc for these pointers

2 Hypothesis testing

Many problems in consulting can be treated as elementary testing problems. First, let's review some of the philosophy of hypothesis testing.

- Testing provides a principled framework for filtering away implausible scientific claims
  - It's a mathematical formalization of Karl Popper's philosophy of falsification
  - Reject the null hypothesis if the data are not consistent with it, where the strength of the discrepancy is formally quantified

- There are two kinds of errors we can make: (1) Accidentally falsify when true (false positive / type I error) and (2) fail to falsify when actually false (false negative / type II error)

For this analysis paradigm to work, a few points are necessary.

- We need to be able to articulate the sampling behavior of the system under the null hypothesis.
- We need to be able to quantitatively measure discrepancies from the null. Ideally we would be able to measure these discrepancies in a way that makes as few errors as possible – this is the motivation behind optimality theory.

While testing is fundamental to much of science, and to a lot of our work as consultants, there are some limitations we should always keep in mind,

- Often, describing the null can be complicated by particular structure present within a problem (e.g., the need to control for values of other variables). This motivates inference through modeling, which is reviewed below.
- Practical significance is not the same as statistical significance. A p-value should never be the final goal of a statistical analysis – they should be used to complement figures / confidence intervals / follow-up analysis[1] that provide a sense of the effect size.

2.1 (One-sample, Two-sample, and Paired) t-tests

If I had to make a bet for which test was used the most on any given day, I'd bet it's the t-test. There are actually several variations, which are used to interrogate different null hypothesis, but the statistic that is used to test the null is similar across scenarios.

- The one-sample t-test is used to measure whether the mean of a sample is far from a preconceived population mean.
- The two-sample t-test is used to measure whether the difference in sample means between two groups is large enough to substantiate a rejection of the null hypothesis that the population means are the same across the two groups.

What needs to be true for these t-tests to be valid?

- Sampling needs to be independent and identically distributed (i.i.d.), and in two-sample setting, the two groups need to be independent. If this is not the case, you can try pairing or developing richer models, see below.

[1] E.g., studying contributions from individual terms in a χ-square test

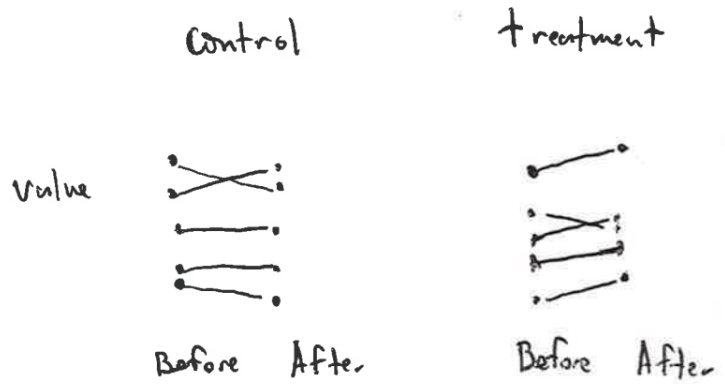

Figure 1: Pairing makes it possible to see the effect of treatment in this toy example. The points represent a value for patients (say, white blood cell count) measured at the beginning and end of an experiment. In general, the treatment leads to increases in counts on a per-person basis. However, the inter-individual variation is very large – looking at the difference between before and after without the lines joining pairs, we wouldn't think there is much of a difference. Pairing makes sure the effect of the treatment is not swamped by the variation between people, by controlling for each persons' white blood cell count at baseline.

- In the two sample case, depending on the the sample sizes and population variances within groups, you would need to use different estimates of the standard error.
- If the sample size is large enough, we *don't* need to assume normality in the population(s) under investigation. This is because the central limit kicks in and makes the means normal. In the small sample setting however, you would need normality of the raw data for the t-test to be appropriate. Otherwise, you should use a nonparametric test, see below.

Pairing is a useful device for making the t-test applicable in a setting where individual level variation would otherwise dominate effects coming from treatment vs. control. See Figure 1 for a toy example of this behavior.

- Instead of testing the difference in means between two groups, test for whether the per-individual differences are centered around zero.
- For example, in Darwin's Rhea Mays data, a treatment and control plant are put in each pot. Since there might be a pot-level effect in the growth of the plants, it's better to look at the per-pot difference (the differences are i.i.d).

Pairing is related to a few other common statistical ideas,

- Difference in differences: In a linear model, you can model the change from baseline
- Blocking: tests are typically more powerful when treatment and control groups are similar to one another. For example, when testing whether two types of soles for shoes have different degrees of wear, it's better to give one of each type for each person in the study (randomizing left vs. right foot) rather than randomizing across people and assigning one of the two sole types to each person

Some examples from past consulting quarters,

- Interrupted time analysis
- Effectiveness of venom vaccine
- The effect of nurse screening on hospital wait time
- Testing the difference of mean in time series
- t-test vs. Mann-Whitney
- Trial comparison for walking and stopping
- Nutrition trends among rural vs. urban populations

2.2 Difference in proportions

2.3 Contingency tables

Contingency tables are a useful technique for studying the relationship between categorical variables. Though it's possible to study K-way contingency tables (relating K categorical variables), we'll focus on 2×2 tables, which relate two categorical variables with two levels each. These can be represented like in the table in Table 2.3. We usually imagine a sampling mechanism that leads to this table [2], where the probability that a sample lands in cell ij is p_{ij}. Hypotheses are then formulated in terms of these p_{ij}.

A few summary statistics of 2×2 tables are referred to across a variety of tests,

- Difference in proportions: This is the difference $p_{12} - p_{22}$. If the columns represent the survival after being given a drug, and the rows correspond to treatment vs. control, then this is the difference in probabilities someone will survive depending on whether they were given the treatment drug or the control / placebo.
- Relative Risk: This is the ratio $\frac{p_{12}}{p_{22}}$. This can be useful because a small difference near zero or near one is more meaningful than a small difference near 0.5.

[2] The most common are binomial, multinomial, or Poisson, depending on whether we condition on row totals, the total count, or nothing, respectively

	A1	A2	total
B1	n_{11}	n_{12}	$n_{1.}$
B2	n_{21}	n_{22}	$n_{2.}$
total	$n_{.1}$	$n_{.2}$	$n_{..}$

Table 1: The basic representation of a 2×2 contingency table.

- Odds-Ratio: This is $\frac{p_{12}p_{21}}{p_{11}p_{22}}$. It's referred to in many test, but I find it useful to transform back to relative risk whenever a result is state in terms of odds ratios.

- A cancer study
- Effectiveness of venom vaccine
- Comparing subcategories and time series
- Family communication of genetic disease

2.3.1 χ^2 tests

The χ^2 test is often used to study whether or not two categorical variables in a contingency table are related. More formally, it assesses the plausibility of the null hypothesis of independence,

$$H_0 : p_{ij} = p_{i+}p_{+j}$$

The two most common statistics used to evaluate discrepancies the Pearson and likelihood ratio χ^2 statistics, which measure the deviation from the expected count under the null,

- Pearson: Look at the squared absolute difference between the observed and expected counts, using $\sum_{i,j} \frac{(n_{ij}-\hat{\mu}_{ij})^2}{\hat{\mu}_{ij}}$
- Likelihood-ratio: Look at the logged relative difference between observed and expected counts, using $2\sum_{i,j} n_{ij} \log\left(\frac{n_{ij}}{\hat{\mu}_{ij}}\right)$

Under the null hypotheses, and assuming large enough sample sizes, these are both χ^2 distributed, with degrees of freedom determined by the number of levels in each categorical variable.

A useful follow-up step when the null is rejected is to see which cell(s) contributed to the most to the χ^2-statistic. These are sometimes called Pearson residuals.

2.3.2 Fisher's Exact test

Fisher's Exact test is an alternative to the χ^2 test that is useful when the counts within the contingency table are small and the χ^2 approximation is not necessarily reliable.

- It tests the same null hypothesis of independence as the χ^2-test
- Under that null, and assuming a binomial sampling mechanism (condition on the row and column totals), the count of the top-left cell can be shown to follow a hypergeometric distribution (and this cell determines counts in all other cells).
- This can be used to determine the probability of seeing tables with as much or more extreme departures from independence.
- There is a generalization to $I \times J$ tables, based on the multiple hypergeometric distribution
- The most famous example used to explain this test is the Lady Tasting Tea.

2.3.3 Cochran-Mantel-Haenzel test

The Cochran-Mantel Haenzel test is a variant of the exact test that applies when samples have been stratified across K groups, yielding K 2×2 separate contingency tables[3].

- The null hypothesis to which this test applies is that, in each of the K strata, there is no association between rows and columns.
- The test statistic consists of pooling deviations from expected counts across all K strata, where the expected counts are defined conditional on the margins (they are the means and variances under a hypergeometric distribution),

$$\frac{\sum_{k=1}^{K} \left(n_{11k} - \mathbb{E}\left[n_{11k}\right]\right)^2}{\sum_{k=1}^{K} \mathrm{Var}\left(n_{11k}\right)}$$

Some related past problems,

- Mantel haenzel chisquare test

[3] These are sometimes called partial tables

2.3.4 McNemar's test

McNemar's test is used to test symmetry in marginal probabilities across the diagonal in a contingency table.

- More formally, the null hypothesis asks whether the running marginal probabilities across rows and columns are the same: $p_{i\cdot} = p_{\cdot i}$ for all i.
- This is the so-called "test of marginal homogeneity," it it is often used to see whether a treatment had any effect. For example, if the rows of the contingency table measure whether someone is sick before the treatment and the columns measure whether they were still sick afterwards, then if the probability that they are sick has not changed between the timepoints, then the treatment has had no effect (and the null hypothesis of marginal homogeneity holds).
- The test statistic[4] used by McNemar's test is given by

$$\frac{n_{21} - n_{12}}{\sqrt{n_{21} + n_{12}}},$$

  which measures the discrepancy in off-diagonal elements.

2.4 Nonparametric tests

There are reasons we may prefer a nonparametric test to a parametric one,

- The assumptions for a particular test statistic to be valid do not hold. This was the motivation for switching from the χ^2-test to Fisher's exact test when the counts in the contingency table are small.
- The test statistic of interest has no known distribution. It is sometimes the case that a test statistic is proposed for a problem, but that its distribution under the null hypothesis has no analytical form.

Related to the first point, we will see several alternatives to the different t-tests when the sample size is so small that the central limit theorem cannot be expected to hold for the sample mean. For the second point, we will consider some general devices for simulating the distribution of a test statistic under its null distribution – these methods typically substitute complicated mathematics with intensive computation.

2.4.1 Rank-based tests

Rank-based tests are often used as a substitute for t-tests in small-sample settings. The most common are the Mann-whitney (substitute for 2-sample t-test), sign (substitute for paired t-test), and signed-rank tests (also a substitute for paired t-test, but using more information than the sign test). Some details are given below.

[4] It's actually just a score statistic, for those who're familiar with that.

- Mann-Whitney test
  - The null hypothesis is

$$H_0 : \mathbb{P}(X > Y) = \mathbb{P}(Y > X),$$

    which is a strictly stronger condition than equality in the means of X and Y. For this reason, care needs to be taken to interpret a rejection in this and other rank-based tests – the rejection could have been due to any difference in the distributions (for example, in the variances), and not just a difference in the means.
  - The procedure does the following: (1) combine the two groups of data into one pooled set, (2) rank the elements in this pooled set, and (3) see whether the ranks in one group are systematically larger than another. If there is such a discrepancy, reject the null hypothesis.
- Sign test
  - This is an alternative to the paired t-test, when data are paired between the two groups (think of a change-from-baseline experiment).
  - The null hypothesis is that the differences between paired measurements is symmetrically distributed around 0.
  - The procedure first computes the sign of the difference between all pairs. It then computes the number of times a positive sign occurs and compares it with the a Bin $\left(n, \frac{1}{2}\right)$, which is how we'd expect this quantity to be distributed under the null hypothesis.
  - Since this test only requires a measurement of the sign of the difference between pairs, it can be applied in settings where there is no numerical data (for example, data in a survey might consist of "likes" and "dislikes" before and after a treatment).
- Signed-rank test
  - In the case that it is possible to measure the size of the difference between pairs (not just their sign), it is often possible to improve the power of the sign test, using the signed-rank test.
  - Instead of simply calculating the sign of the difference between pairs, we compute provide a measure of the size of the difference between pairs. For example, in numerical data, we could just use $|x_{i\text{after}} - x_{i\text{before}}|$.
  - At this point, order the difference scores from largest to smallest, and see whether one group is systematically overrepresented among the larger scores[5]. In this case, reject the null hypothesis.
- Evaluating results of a training program

[5] The threshold is typically tabulated, or a more generally applicable normal approximation can be applied

- Nonparametric tests for mean / variance
- t-test vs. Mann-Whitney
- Trial comparison for walking and stopping

2.4.2 Permutation tests

Permutation tests are a kind of computationally intensive test that can be used quite generally. The typical setting in which it applies has two groups between which we believe there is some difference. The way we measure this difference might be more complicated than a simple difference in means, so no closed-form distribution under the null may be available.

The basic idea of the permutation test is that we can randomly create artificial groups in the data, so there will be no systematic differences between the groups. Then, computing the statistic on these many artificial sets of data gives us an approximation to the null distribution of that statistic. Comparing the value of that statistic on the observed data with this approximate null can give us a p-value. See Figure 2 for a representation of this idea.

- More formally, the null hypothesis tested by permutation tests is that the group labels are *exchangeable* in the formal statistical sense[6]
- For the same reason that caution needs to be exercised when interpreting rejections in the Mann-Whitney test, it's important to be aware that a permutation test can reject the null for reasons other than a simple difference in means.
- The statistic you use in a permutation test can be whatever you want, and the test will be valid. Of course, the power of the test will depend crucially on whether the statistic is tailored to the type of departures from the null which actually exist.
- The permutation p-value of a test statistic is obtained by making a histogram of the statistic under all the different relabelings, placing the observed value of the statistic on that histogram, and looking at the fraction of the histogram which is more extreme than the value calculated from the real data. See Figure 2.
- A famous application of this method is to Darwin's Zea Mays data[7]. In this experiment, Darwin planted Zea Mays that had been treated in two different ways (self vs. cross-fertilized). In each pot, he planted two of each plant, and he made sure to put one of each type in each pot, to control for potential pot-level effects. He then looked to see how high the plants grew. The test statistic was then the standardized difference in means, and this was computed many times after randomly relabeling the

[6] The distribution is invariant under permutations.

[7] R.A. Fisher also used this dataset to explain the paired t-test.

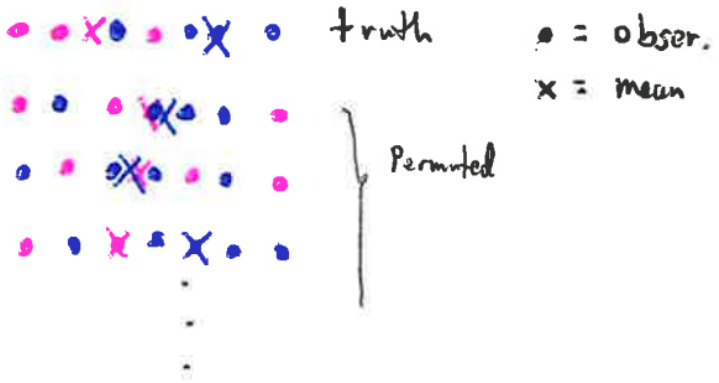

Figure 2: A representation of a two-sample difference in means permutation test. The values along the x-axis represent the measured data, and the colors represent two groups. The two row gives the values in the observed data, while each following row represents a permutation in the group labels of that same data. The crosses are the averages within those groups. Here, it looks like in the real data the blue group has a larger mean than the pink group. This is reflected in the fact that the difference in means in this observed data is larger here than in the permuted data. The proportion of times that the permuted data have a larger difference in means is used as the p-value.

plants as self and cross-fertilized. The actual difference in heights for the groups was then compared to this histogram, and the difference was found to be larger than those in the approximate null, so the null hypothesis was rejected.

Some examples of permutation tests recommended in past consulting quarters,

- Diference of two multinomials
- Comparing subcategories and time series
- Changepoint in time course of animal behavior

2.4.3 Bootstrap tests

While the bootstrap is typically used to construct confidence intervals (see Section 3.2), it is also possible to use the bootstrap principle to perform hypothesis test. Like permutation tests, it can be applied in a range of situations where classical testing may not be appropriate.

- The main idea is to simulate data under the null and calculate the test statistic on these null data sets. The p-value can then be calculated by

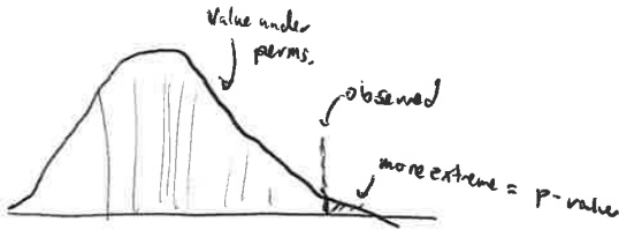

Figure 3: To compute a p-value for a permutation test, refer to the permutation null distribution. Here, the histogram provides the value of the test statistic under many permutations of the group labeling – this approximates how the test statistic is distributed under the null hypothesis. The value of the test statistic in the observed data is the vertical bar. The fraction of the area of this histogram that has a more extreme value of this statistic is the p-value, and it exactly corresponds to the usual interpretation of p-values as the probability under the null that I observe a test statistic that is as or more extreme.

making a comparison of the real data to this approximate null distribution, as in permutation tests.

- As in permutation tests, this procedure will always be valid, but will only be powerful if the test-statistic is attuned to the actual structure of departures from the null.
- The trickiest part of this approach is typically describing an appropriate scheme for sampling under the null. This means we need to estimate $\hat{F}_0$ among a class of CDFs $\mathcal{F}$, consistent with the null hypothesis.
- For example, in a two-sample difference of means testing situation, to sample from the null, we center each group by subtracting away the mean, so that H_0 actually holds, and then we simulate new data by sampling with replacement from this pooled, centered histogram.

2.4.4 Kolmogorov-Smirnov

The Kolmogorov-Smirnov (KS) test is a test for either (1) comparing two groups of real-valued measurements or (2) evaluating the goodness-of-fit of a collection of real-valued data to a prespecified reference distribution.

- In its two-sample variant, the empirical CDFs (ECFs) for each group are calculated. The discrepancy is measured by the largest absolute gap between the two ECDFs. This is visually represented in Figure 4.
- The distribution of this gap under the null hypothesis that the two groups have the same ECDF was calculated using an asymptotic approximation, and this is used to provide p-values.

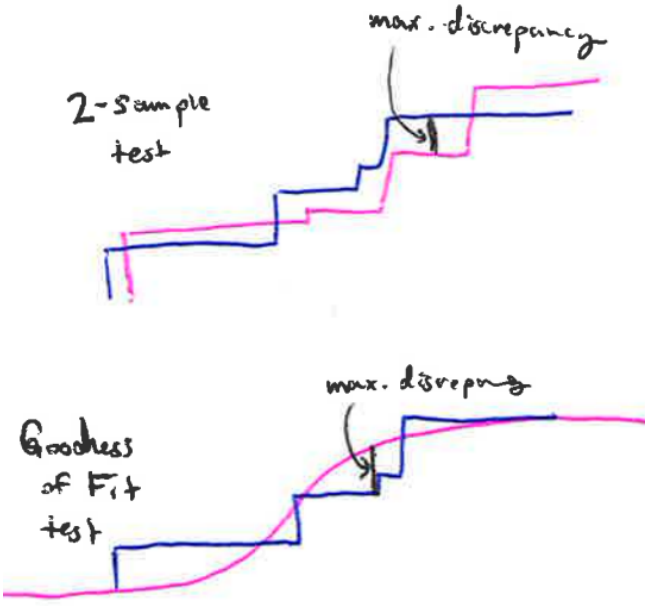

Figure 4: The motivating idea of the two-sample and goodness-of-fit variants of the KS test. In the 2-sample test variant, the two colors represent two different empirical CDFs. The largest vertical gap between these CDFs is labeled by the black bar, and this defines the KS statistic. Under the null that the two groups have the same CDF, this statistic has a known distribution, which is used in the test. In the goodness-of-fit variant, the pink line now represents the true CDF for the reference population. This test sees whether the observed empirical CDF (blue line) is consistent with this reference CDF, again by measuring the largest gap between the pair.

- In the goodness-of-fit variant, all that changes is that one of the ECDFs is replaced with the known CDF for the reference distribution.

- Reference for KS test

2.5 Power analysis

Before performing an experiment, it is important to get a rough sense of how many samples will need to be collected in order for the follow-up analysis to have a chance at detecting phenomena of interest. This general exercise is called a power-analysis, and it often comes up in consulting sessions because many grant agencies will require a power analysis be conducted before agreeing to provide funding.

2.5.1 Analytical

Traditionally, power analysis have been done by deciding in advance upon the type of statistical test to apply to the collected data and then using basic statistical theory to work out exactly the number of samples required to reject the null when the signal has some assumed strength.

- For example, if the true data distribution is assumed to be $\mathcal{N}\left(\mu, \sigma^2\right)$, and we are testing against the null $\mathcal{N}\left(0, \sigma^2\right)$ using a one-sample t-test, then the fact that $\bar{x} \sim \mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ can be used to analytically calculate the probability that the observed mean will be above the t-test rejection threshold.
- The size of the signal is assumed known (smaller signals require larger sample sizes to detect). Of course this is the quantity of interest in the study – if it were known, there would be no point in doing the study. The idea though is to get a rough estimate of the number of samples required for a few different signal strengths[8].
- There are many power calculators available, these can be useful to share / walk through with clients.

2.5.2 Computational

When more complex tests or designs are used, it is typically impossible to work out an analytical form for the sample size as a function of signal strength. In this situation, it is common to set up a simulation experiment to approximate this function.

- The client needs to specify a simulation mechanism under which the data can be plausibly generated, along with a description of which knobs change the signal strengths in what ways.
- The client needs to specify the actual analysis that will be applied to these data to declare significance.
- From here, many simulated datasets are generated for every configuration of signal strengths along with a grid of sample sizes. The number of times the signal was correctly detected is recorded and is used to estimate of the power under each configuration of signal strength and sample size.
- Appropriate sample size calculations
- Land use and forest cover
- t-test vs. Mann-Whitney

[8] Sometimes, a pilot study has been conducted previously, which can give an approximate range for the signal strength to expect

3 Elementary estimation

While testing declares that a parameter θ cannot plausibly lie in the subset $\mathcal{H}_0$ of parameters for which the null is true, estimation provides an estimate $\hat{\theta}$ of θ, typically accompanied by a confidence interval, which summarizes the precision of that estimate.

In most multivariate situations, estimation requires the technology of linear modeling. However, in simpler settings, it is often possible to construct a confidence interval based parametric theory or the bootstrap.

3.1 Classical confidence intervals

Classical confidence intervals are based on rich parametric theory. Formally, a confidence interval for a parameter θ is random (data dependent) interval $[L(X), U(X)]$ such that, under data X generated with parameter θ, that interval would contain θ some prespecified percentage of the time (usually 90, 95, or 99%).

- The most common confidence interval is based on the Central Limit Theorem. Suppose data $x_1, \ldots, x_n$ are sampled i.i.d. from a distribution with mean θ and variance σ^2. Then the fact that $\sqrt{n}(\bar{x}_n - \theta) \approx \mathcal{N}(0, \sigma^2)$ for large n means that $\left[\bar{x}_n - z_{\alpha/2}\frac{\sigma}{\sqrt{n}}, \bar{x}_n + z_{1-\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right]$, where $z_{\frac{\alpha}{2}}$ is the $\frac{\alpha}{2}$ quantile of the normal distribution, is a $(1-\alpha)\,\%$ confidence interval for θ.

- Since proportions can be thought of as averages of indicators variables (1 if present, 0 if not) which have bernoulli means p and variances $p(1-p)$, the same reasoning gives confidence intervals for proportions.

- For the same reason that we might prefer a t-test to a z-test[9], we may sometimes prefer using a t-quantile instead.

- If a confidence interval is known for a parameter, then it's easy to construct an approximate interval for any smooth function of that parameter using the delta method. For example, this is commonly used to calculate confidence intervals for log-odds ratios.

3.2 Bootstrap confidence intervals

There are situations for which the central limit theorem might not apply and no theoretical analysis can provide a valid confidence interval. For example, we might have defined a parameter of interest that is not a simple function of means of the data. In many of these cases, it may be nonetheless be possible to use the bootstrap.

[9] Samples sizes too small to put faith in the central limit theorem

- The main idea of the bootstrap is the "plug-in principle." Suppose our goal is to calculate the variance of some statistic $t(X_{1:n})$ under the true sampling distribution F for the X_i. That is, we want to know[10] $\text{Var}_F\left(\hat{\theta}(X_{1:n})\right)$, but this is unknown, since we don't actually know F. However, we can plug-in $\hat{F}$ for F, which gives $\text{Var}_{\hat{F}}\left(\hat{\theta}(X_{1:n})\right)$. This two is unknown, *but* we can sample from $\hat{F}$ to approximate it – the more samples from $\hat{F}$ we make, the better our estimate of $\text{Var}_{\hat{F}}\left(\hat{\theta}(X_{1:n})\right)$. This pair of approximations (plugging in $\hat{F}$ for F, and then simulating to approximate the variance under $\hat{F}$) gives a usable approximation $\hat{v}\left(\hat{\theta}\right)$ of $\text{Var}_F\left(\hat{\theta}(X_{1:n})\right)$. The square-root of this quantity is usually called the bootstrap estimate of standard error.

- The bootstrap estimate of standard error can be used to construct confidence intervals,

$$\left[\hat{\theta}(X_{1:n}) - z_{\frac{\alpha}{2}}\sqrt{\frac{\hat{v}\left(\hat{\theta}\right)}{n}}, \hat{\theta}(X_{1:n}) + z_{1-\frac{\alpha}{2}}\sqrt{\frac{\hat{v}\left(\hat{\theta}\right)}{n}}\right]$$

- Since sampling from $\hat{F}$ is the same as sampling from the original data with replacement, the bootstrap is often explained in terms of resampling the original data.

- A variant of the above procedure skips the calculation of a variance estimator and instead simply reports the upper and lower α percentiles of the samples of $\hat{\theta}(X^*_{1:n})$ for $X^*_i \sim \hat{F}$. This is sometimes called the percentile bootstrap, and it is the one more commonly encountered in practice.

- In consulting situations, the bootstrap gives very general flexibility in defining statistics on which to do inference – you can do inference on parameters that might be motivated by specific statistical structure or domain knowledge.

4 (Generalized) Linear Models

Linear models provide the basis for most inference in multivariate settings. We won't even begin to try to cover this topic comprehensively – there are entire course sequences that only cover linear models. But, we'll try to highlight the main regression-related ideas that are useful to know during consulting.

This section is focused more on the big-picture of linear regression and when we might want to use it in consulting. We defer a discussion of inference in linear models to Section 5.

[10] We usually want this so we can calculate a confidence interval, see the next bullet.

Some past consulting problems:

- GLM with bounds on response

4.1 Linear regression

Linear regression learns a linear function between covariates and a response, and is popular because there are well-established methods for performing inference for common hypothesis.

- Generally, model-fitting procedures suppose that there is a single variable Y of special interest. The goal of a supervised analysis is to determine the relationship between many other variables (called covariates), $X_1, \ldots, X_p$ and Y. Having a model $Y = f(X_1, \ldots, X_p)$ can be useful for many reasons, especially (1) improved scientific understanding (the functional form of f is meaningful) and (2) prediction, using f learned on one set of data to guess the value of Y on a new collection of X's.

- Linear models posit that the functional form f is *linear* in the $X_1, dots, X_p$. This is interpreted geometrically by saying that the change in Y that occurs when X_j is changed by some amount (and all other covariates are held fixed) is proportional to the change in X_j. E.g., when $p = 1$, this means the scatterplot of X_1 against Y can be well-summarized by a line.

- A little more formally, we suppose we have observed samples indexed by i, where $x_i \in \mathbb{R}^p$ collects the covariates for the i^{th} sample, and y_i is the associated observed response. A regression model tries to find a parameter $\beta \in \mathbb{R}^p$ so that

$$y_i = x_i^T \beta + \epsilon_i \tag{1}$$

  is plausible, where ϵ_i are drawn i.i.d. from $\mathcal{N}\left(0, \sigma^2\right)$ for some (usually unknown) σ^2. The fitted value for β after running a linear regression is denoted $\hat{\beta}$.

- Compared to other forms of model fitting, a major advantage of of linear models is that inference is usually relatively straightforwards – we can do tests of significance / build confidence intervals of the strength of association across different X's as well as comparison between models with different sets of variables.

- The linear assumption is not well-suited to binary, count, or categorical responses Y, because it the model might think the response Y belongs to some range that's not even possible (think of extrapolating a linear in a scatterplot when the y-axis values are all between 0 and 1). In these situations, it is necessary to apply generalized linear models (GLMs) (GLMs). Fortunately, many of the ideas of linear models (methods for inference in particular) have direct analogs in the GLM setting.

When is linear regression useful in consulting?

- In a consulting setting, regression is useful for understanding the association between two variables, controlling for many others. This is basically a rephrasing of point (2) above, but it's the essential interpretation of linear regression coefficients, and it's this interpretation that many research studies are going after.
- Sometimes a client might originally come with a testing problem, but might want help extending it to account for additional structure or covariates. In this setting, it can often be useful to propose a linear model instead: it still allows inference, but it becomes much easier to encode more complex structure.

What are some common regression tricks useful in consulting?

- Adding interactions: People will often ask about adding interactions in their regression, but usually from an intuition about the non-quantitative meaning of the word "interaction." It's important to clarify the quantitative meaning: Including an interaction term between X_1 and X_2 in a regression of these variables onto Y means that the slope of the relationship between X_1 on Y will be different depending on the value of X_2. For example, if X_2 can only take on two values (say, A and B), then the relationship between X_1 and Y will be linear with slope β_{1A} in the case that X_2 is A and β_{1B} otherwise[11]. When X_2 is continuous, then there is a continuum of slopes depending on the value of X_2: $\beta_1 + \beta_{1\times 2}X_2$. See Figure 5 for a visual interpretation of interactions.
- Introducing basis functions: The linearity assumption is not as restrictive as it seems, if you can cleverly apply basis functions. Basis functions are functions like polynomials (or splines, or wavelets, or trees...) which you can mix together to approximate more complicated functions, see Figure 6. Linear mixing can be done with linear models.

  To see why this is potentially useful, suppose we want to use time as a predictor in a model (e.g., where Y is number of species j present in the sample), but that species population doesn't just increase or decrease linearly over time (instead, it's some smooth curve). Here, you can introduce a spline basis associated with time and then use a linear regression of the response onto these basis functions. The fitted coefficients will define a mean function relating time and the response.
- Derived features: Related to the construction of basis functions, it's often possible to enrich a linear model by deriving new features that you imagine might be related to Y. The fact that you can do regression onto variables

[11] In terms of the regression coefficients β_1 for the main effect of X_1 on Y and $\beta_{1\times 2}$ for the interaction between X_1 and X_2, this is expressed as $\beta_{1A} = \beta_1$ and $\beta_{1B} = \beta_1 + \beta_{1\times 2}$.

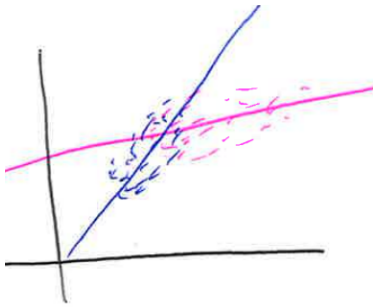

Figure 5: In the simplest setting, an interaction between a continuous and binary variable leads to two different slopes for the continuous variable. Here, we are showing the scatterplot (x_{i1}, y_i) pairs observed in the data. We suppose there is a binary variable that has also been measured, denoted x_{i2}, and we shade in each point according to its value for this binary variable. Apparently, the relationship between x_1 and y depends on the value of y (when in the pink group, the slope is less. This can exactly be captured by introducing an interaction term between x_1 and x_2. In cases where x_2 is not binary, we would have a continuous of slopes between x_1 and y – one for each value of x_2.

that aren't just the ones that were collected originally might not be obvious to your client. For example, if you were trying to predict whether someone will have a disease[12] based on time series of some lab tests, you can construct new variables corresponding to the "slope at the beginning," or "slope at the end" or max, or min, ... across the time series. Of course, deciding which variables might actually be relevant for the regression will depend on domain knowledge.

- One trick – introducing random effects – is so common that it gets it's own section. Basically, it's useful whenever you have a lot of levels for a particular categorical vector.

Some examples where regression was used in past sessions,

- Family communication genetic disease
- Stereotype threat in virtual reality
- Fish gonad regression
- UV exposure and birth weight
- Land use and forest cover
- Molecular & cellular physiology

[12] In this case, the response is binary, so you would probably use logistic regression, but the basic idea of derived variables should still be clear.

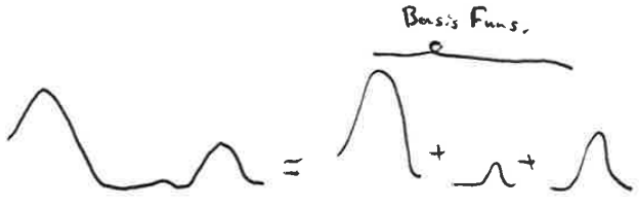

Figure 6: Complex functions can be represented as simple mixtures of basis functions. This allows the modeling of nonlinear effects using just a linear model, by including the basis functions among the set of covariates.

- Multiple linear regression for industry data
- Prediction intervals

4.2 Diagnostics

How can you assess whether a linear regression model is appropriate? Many types of diagnostics have been proposed, but a few of the most common are,

- Look for structure in residuals: According to equation 1, the amount that the predictions $\hat{y}_i = x_i^T \hat{\beta}$ is off from y_i (this difference is called the residual $r_i = \hat{y}_i - y_i$) should be about i.i.d. $\mathcal{N}\left(0, \sigma^2\right)$. Whenever there is a systematic pattern in these residuals, the model is misspecified in some way. For example, if you plot the residuals over time and you find clumps that are all positive or negative, it means there is some unmeasured phenomena associated with these time intervals that influences the average value of the responses. In this case, you would define new variables for whether you are in one of these intervals, but the solution differs on a case-by-case basis. Other types of patterns to keep an eye out for: nonconstant spread (heteroskesdasticity), large outliers, any kind of discreteness (see Figure 7).
- Make a qq-plot of residuals: More generally than simply finding large outliers in the residuals, we might ask whether the residuals are plausibly drawn from a normal distribution. qq-plots give one way of doing this – more often than not the tails are heavier than normal. Most people ignore this, but it can be beneficial to consider e.g. robust regression or considering logistically (instead of normally) distributed errors.
- Calculate the leverage of different points: The leverage of a sample is a measure of how much the overall fit would change if you took that point out. Points with very high leverage can be cause for concern – it's bad if your fit completely depended on one or two observations only – and these high leverage points often turn out to be outliers. See Figure 8 for an example of this phenomenon. If you find a few points have very high

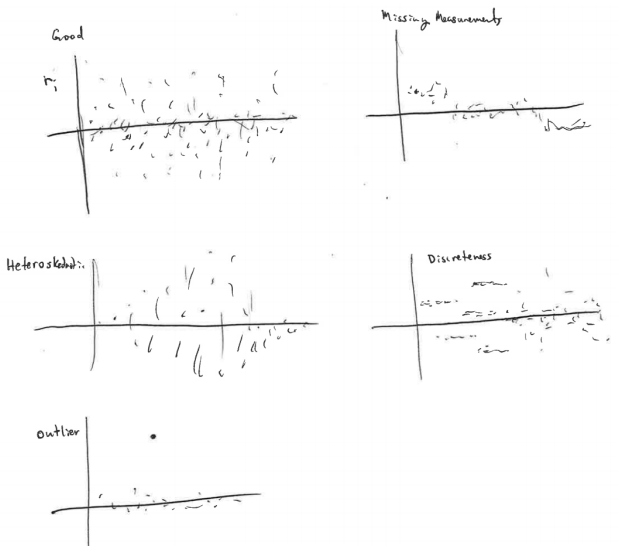

Figure 7: Some of the most common types of "structure" to watch out for in residual plots are displayed here. The top left shows how residuals should appear – they look essentially i.i.d. $\mathcal{N}(0,1)$. In the panel below, there is nonconstant variance in the residuals, the one on the bottom has an extreme outlier. The panel on the top-right seems to have means that are far from zero in a structured way, while the one below has some anomalous discreteness.

leverage, you might consider throwing them out. Alternatively, you could consider a robust regression method.

- Simulate data from the model: This is a common practice in the Bayesian community ("posterior predictive checks"), though I'm not aware if people do this much for ordinary regression models. The idea is simple though – simulate data from $x_i^T \hat{\beta} + \epsilon_i$, where $\epsilon_i \sim \mathcal{N}\left(0, \hat{\sigma}^2\right)$ and see whether the new y_i's look comparable to the original y_i's. Characterizing the ways they don't match can be useful for modifying the model to better fit the data.

Some diagnostics-related questions from past quarters,

- Evaluating regression model

4.3 Logistic regression

Logistic regression is the analog of linear regression that can be used whenever the response Y is binary (e.g., patient got better, respondent answered "yes," email was spam).

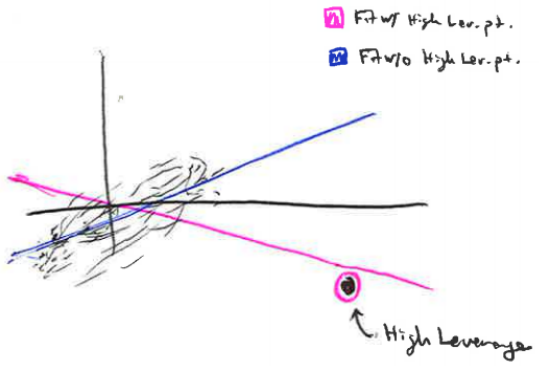

Figure 8: The leverage of a sample can be thought of as the amount of influence it has in a fit. Here, we show a scatterplot onto which we fit a regression line. The cloud near the origin and the one point in the bottom right represent the observed samples. The blue line is the regression fit when ignoring the point on the bottom right, while the pink line is the regression including that point. Evidently, this point in the bottom right has very high leverage – in fact, it reverses the sign of the association between X and Y. This is also an example of how outliers (especially outliers in the X direction) can have very high leverage.

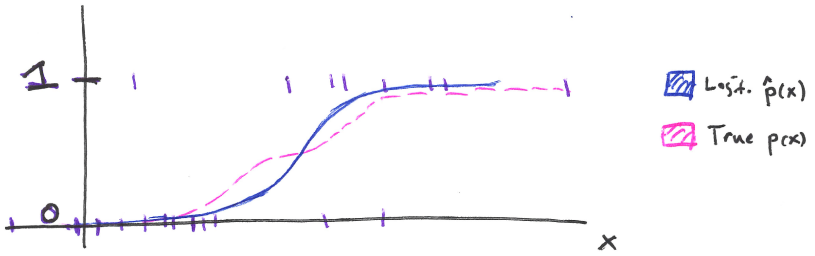

Figure 9: An example of the type of approximation that logistic regression makes. The x-axis represent the value of the feature, and the y-axis encodes the binary 0 / 1 response. The purple marks are observed (x_i, y_i) pairs. Note that class 1 becomes more common when x is large. The pink line represents the "true" underlying class probabilities as a function of x, which we denote as $p(x)$. This doesn't lie in the logistic family $\frac{1}{1+\exp(-x\beta)}$, but it can be approximated by a member of that family, which is drawn in blue (this is the logistic regression fit).

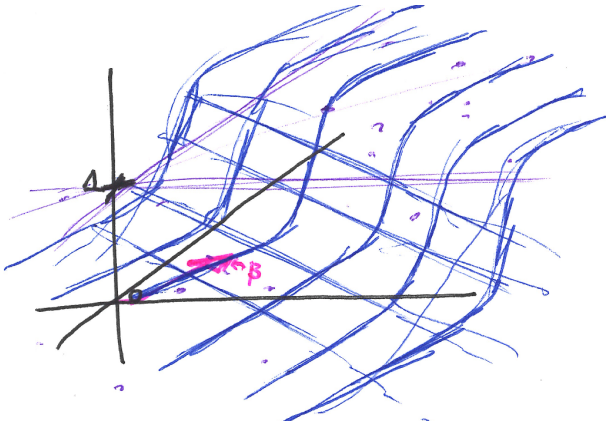

Figure 10: Figure 9 can be generalized to higher dimensions. The vertical axis still represents the class label, but the horizontal plane now encodes values for two variables, x_1 and x_2. The fitted logistic regression surface is drawn in blue – we can see that the probability of class 1 increases when both x_1 and x_2 are large. β now controls the orientation of this logistic regression surface.

- In linear regression, the response y are directly used in a model of the form $y_i = x_i^T \beta + \epsilon_i$. In logistic regression, we now want a model between the x_i and the unknown probabilities of the two classes when we're at x_i: $p(x_i)$ and $1 - p(x_i)$.
- The observed value of y_i corresponding to x_i is modeled as being drawn from a coin flip with probability $p(x_i)$.
- If we had fit an ordinary linear regression model to the y_i, we might get fitted responses $\hat{y}_i$ outside of the valid range $[0, 1]$, which in addition to being confusing is bad modeling.
- Logistic regression instead models the log-odds transformation to the $p(x_i)$ respectively). Concretely, it assumes the model

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p_\beta(x_i)^{\mathbb{I}(y_i=1)} (1 - p_\beta(x_i))^{\mathbb{I}(y_i=0)} \tag{2}$$

  where we are approximating

$$p(x_i) \approx p_\beta(x_i) := \frac{1}{1 + \exp\left(-x_i^T \beta\right)} \tag{3}$$

  Logistic regression fits the parameter β to maximize the likelihood defined in the model .
- An equivalent reformulation of the assumption is that $\log \frac{p(x_i)}{1-p(x_i)} \approx x_i^T \beta$, i.e. the log-odds are approximately linear.

- Out of the box, the coefficients β fitted by logistic regression can be difficult to interpret. Perhaps the easiest way to interpret them is in terms of the relative risk, which gives an analog to the usual linear regression interpret "when the j^{th} feature goes up by one unit, the expected response goes up by β_j." First, recall that the relative risk is defined as

$$\frac{\mathbb{P}\left(y_i = 1 | x_i\right)}{\mathbb{P}\left(y_i = 0 | x_i\right)} = \frac{p\left(x_i\right)}{1 - p\left(x_i\right)}, \tag{4}$$

  which in logistic regression is approximated by $\exp\left(x_i^T \beta\right)$. If we increase the j^{th} coordinate of x_i (i.e., we take $x_i \to x_i + \delta_j$), then this relative risk becomes

$$\exp\left(\left(x_i + \delta_j\right)^T \beta\right) = \exp\left(x_i^T \beta\right) \exp\left(\beta_j\right). \tag{5}$$

  The interpretation is that the relative risk got multiplied by $\exp\left(\beta_j\right)$ when we increased the j^{th} covariate by one unit.

- Diagnostics: While you could study the differences $y_i - \hat{p}\left(x_i\right)$, which would be analogous to linear regression residuals, it is usually more informative to study the Pearson or deviance residuals, which upweight small differences near the boundaries 0 and 1. These types of residuals take into account structure in the assumed bernoulli (coin-flipping) sampling model.

- For ways to evaluate the classification accuracy of logistic regression models, see Section 10.5, and for an overview of formal inference, see Section 5.

- Teacher data and logistic regression

- Conditional logistic regression

- Land investment and logistic regression

- Uganda

4.4 Poisson regression

Poisson regression is a type of generalized linear model that is often applied when the responses y_i are counts (i.e., $y_i \in \{0, 1, 2, \ldots, \}$).

- As in logistic regression, one motivation for using this model is that using ordinary logistic regression on these responses might yield predictions that are impossible (e.g., numbers below zero, or which are not integers).

- To see where the main idea for this model comes from, recall that the Poisson distribution with rate λ draws integers with probabilities

$$\mathbb{P}_\lambda\left[y = k\right] = \frac{\lambda^k \exp\left(-\lambda\right)}{k!}. \tag{6}$$

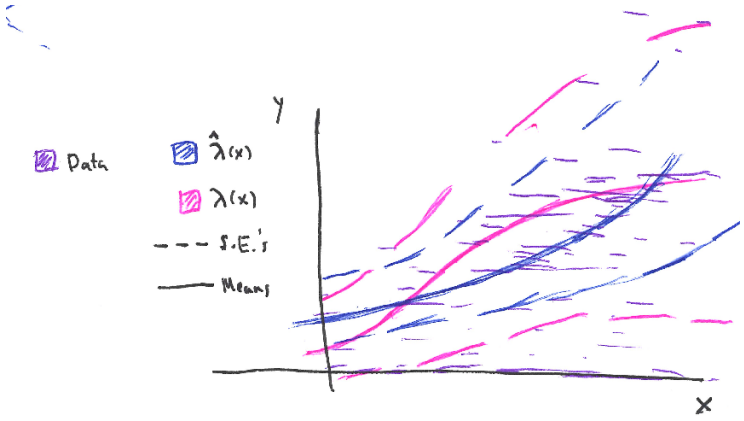

Figure 11: A representation of Poisson regression with one feature. The x-axis encodes the value of this feature, and the y-axis gives the count value. Observed data points are sketched in purple. The mean count $\lambda(x)$ increases as a function of x – the true mean is sketched in pink. Poisson regression models this mean count as an exponential in x, and the approximation is drawn in blue. The dashed lines represent the variation of the counts around there means. Note that larger counts are associated with larger variation, but that the Poisson regression underestimates the true variation – this data seem overdispersed.

The idea of Poisson regression is to say that the different y_i are drawn Poissons with rates that depend on the associated x_i.

- In more detail, we assume that the data have a joint likelihood

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} \frac{\lambda(x_i)^{y_i} \exp(-\lambda(x_i))}{y_i!} \tag{7}$$

and that the log of the rates are linear in the covariates,

$$\log \lambda(x_i) \approx x_i^T \beta. \tag{8}$$

(modeling the logs as linear makes sure that the actual rates are always nonnegative, which is a requirement for the Poisson distribution).

- We think of different regions of the covariate space as having more or less counts on average, depending on this function $\lambda(x_i)$. Moving from x_i to $x_i + \delta_j$ multiplies the rate from $\lambda(x_i)$ to $\exp(\beta_j)\lambda(x_i)$.

- As in logistic regression, it makes more sense to consider the deviance residuals when performing diagnostics.

- A deficiency in Poisson regression models (which often motivates clients to show up to consulting) is that real data often exhibit *overdipersion*

with respect the assumed Poisson model. The issue is that the mean and variance of counts in a Poisson model are tied together: if you sample from then the mean and variance of the associated counts are both λ. In real data, the variance is often larger than the mean, so while the Poisson regression model might do a good job approximating the mean $\lambda(x_i)$ at x_i, the observed variance of the y_i near x_i might be much larger than $\lambda(x_i)$. This motivates the methods in Section .

4.5 Psueo-Poisson and Negative Binomial regression

Pseudo-Poisson and negative binomial regression are two common strategies for addressing overdispersion in count data.

In the pseudo-Poisson setup, a new parameter φ is introduced that sets the relative scale of the variance in comparison to the mean: $\mathrm{Var}(y) = \varphi \mathbb{E}[y]$. This is not associated with any real probability distribution, and the associated likelihood is called a pseudolikelihood. However, φ can be optimized along with β from the usual Poisson regression setup to provide a maximum pseudolikelihood estimate.

In negative binomial regression, the Poisson likelihood is abandoned altogether in favor of the negative binomial likelihood. Recall that the negative binomial (like the Poisson) is a distribution on nonnegative counts $\{0, 1, 2, \ldots, \}$. It has two parameters, p and r,

$$\mathbb{P}_{p,r}[y = k] = \binom{k + r - 1}{k} p^k (1 - p)^r \tag{9}$$

which can be interpreted as the number of heads that appeared before seeing r tails, when flipping a coin with probability p of heads. More important than the specific form of the distribution is the fact that it has two parameters, which allow different variances even for the same mean,

$$\mathbb{E}_{p,r}[y] = \frac{pr}{1 - p} \tag{10}$$

$$\mathrm{Var}_{p,r}(y) = \frac{pr}{(1 - p)^2} = \mathbb{E}_{p,r}[y] + \frac{1}{r}\left(\mathbb{E}_{p,r}[y]\right)^2. \tag{11}$$

In particular, for small r, the variance is much larger than the mean, while for large r, the variance is about equal to the mean (it reduces to the Poisson).

For negative binomial regression, this likelihood 9 is substituted for the Poisson when doing regression, and the mean is allowed to depend on covariates. On the other hand, while the variance is no longer fixed to the mean, it must be the same across all data points. This likelihood model is not exactly a GLM (the negative binomial is not in the exponential family), but various methods for fitting it are available.

There is a connection between the negative binomial and the Poisson that both illuminates potential sources of overdispersion and suggests new algorithms

for fitting overdispersed data: the negative binomial can be interpreted as a "gamma mixture of Poissons." More specifically, in the hierarchical model,

$$y|\lambda \sim \text{Poi}(\lambda) \tag{12}$$

$$\lambda \sim \text{Gam}\left(r, \frac{p}{1-p}\right), \tag{13}$$

which draws a Poisson with a randomly chosen mean parameter, the marginal distribution of y is exactly $\text{NegBin}(r, p)$. This suggests that one reason overdispersed data might arise is that they are actually a mixture of true Poisson subpopulations, each with different mean parameters. This is also the starting point for Bayesian inference of overdispersed counts, which fit Poisson and Gamma distributions according to the hierarchical model 12.

4.6 Loglinear models

Loglinear models give a likelihood-based approach for studying contingency tables. They are especially useful when there are complex hypotheses about the structure of counts within the table or when there are many cross-tabulations $(I \times J \times K \times \dots)$.

- The main idea is to model the ij^{th} cell in a contingency table by a Poisson with mean μ_{ij}. Enforcing different constraints on the relationships between μ_{ij} across cells allows the study of (and inference about) different kinds of structure.

- For example, if the rows (X) and columns (Y) are independent, the mean decomposes as

  $$\mu_{ij} = \mu p(x = i)\, p(y = j) \tag{14}$$

  or equivalently,

  $$\log \mu_{ij} = \lambda + \lambda_i^X + \lambda_j^Y, \tag{15}$$

  where we have defined the λ's as the logs of the corresponding terms in 14. In particular, the log of the mean is linear in the (logged) marginal probabilities of rows and columns, which is where the name loglinear comes from.

- You can interpret the parameters $\lambda_{i\cdot}$ and $\lambda_{\cdot j}$ in terms of relative risks. For example, in the independence model,

  $$\frac{\mu_{i1}}{\mu_{i2}} = \exp\left(\lambda_1^Y - \lambda_2^Y\right). \tag{16}$$

- It's possible to encode more complex structure through these loglinear models. For example, suppose we have a three-way contingency table,

cross-tabulating X, Y, and Z, and we believe that X and Y should be independent, conditional on Z. This means the joint probabilities decompose as

$$p(x,y,z) = p(x,y|z)\,p(z) \tag{17}$$
$$= p(x|z)\,p(y|z)\,p(z) \tag{18}$$
$$= \frac{p(x,z)}{p(z)}\frac{p(x,z)}{p(z)}p(z) \tag{19}$$
$$= \frac{p(x,z)\,p(y,z)}{p(z)} \tag{20}$$

which leads to the loglinear model,

$$\log \mu_{ijk} = \lambda + \lambda_i^X + \lambda_j^Y + \lambda_k^Z + \lambda_{ik}^{XZ} + \lambda_{jk}^{YZ}. \tag{21}$$

The point is that we can directly do inference about parameters in this conditional independence model, without having to use any specialized contingency table tests, and we can modify the model according to structure in the data and reapply the same machinery for inference (no need for new tests).

4.7 Multinomial regression

Multinomial regression is a generalization of logistic regression for when the response can take on one of K categories (not just $K = 2$).

- Here, we want to study the way the probabilities for each of the K classes varies as x varies: $p(y_i = k|x_i)$ for each $k = 1, \ldots, K$. Think of the responses y_i as observations from a K-sided dice, and that different faces are more probable depending on the associated features x_i.
- The approximation 4.3 is replaced with

$$p(y_i = k|x_i) \approx p_W(y_i = k|x_i) := \frac{\exp\left(w_k^T x_i\right)}{\sum_{k'} \exp\left(w_{k'}^T x_i\right)}, \tag{22}$$

  where the parameters $w_1, \ldots, w_K$ govern the relationship between x_i and the probabilities for different classes.
- As is, this model is not identifiable (you can increase one of the w_ks and decrease another, and end up with the exact same probabilities $p_W(y_i = k|x_i)$. To resolve this, one of the classes (say the K^{th}, this is usually chosen to be the most common class) is chosen as a baseline, and we set $w_K = 0$.
- Then the w_ks can be interpreted in terms of how a change in x_i changes the probability of observing k relative to the baseline K. That is, suppose

we increase the j^{th} variable by one unit, so that $x_i \to x_i + \delta_j$. Then, the relative probability against class K changes according to

$$\frac{\frac{p_W(y_i=k|x_i+\delta_j)}{p_W(y_i=K|x_i+\delta_j)}}{\frac{p_W(y_i=k|x_i)}{p_W(y_i=K|x_i)}} = \frac{\exp\left(w_k^T\left(x_i+\delta_j\right)\right)}{\sum_k' \exp\left(\cdots\right)} \frac{\sum_{k'} \exp\left(\cdots\right)}{\exp\left(w_K^T\left(x_i+\delta_j\right)\right)} \tag{23}$$

$$= \exp\left(w_k^T x_i + \delta_j^{w_k}\right) \tag{24}$$

$$= \exp\left(w_k^T x_i\right) \exp\left(w_{kj}\right), \tag{25}$$

where in the second line we used the fact that w_K is a priori constrained to be 0. So, the K-class analog of relative risk for the k^{th} class is multipled by $\exp\left(w_{kj}\right)$ when we increase the j^{th} feature by a unit.

4.8 Ordinal regression

Sometimes we have K classes for the responses, but there is a natural ordering between them. For example, survey respondents might have chosen one of 6 values along a likert scale. Multinomial regression is unaware of this additional ordering structure – a reasonable alternative in this setting is ordinal regression.

- The basic idea for ordinal regression is to introduce a continuous latent variable z_i along with $K-1$ "cutpoints" $\gamma_1, \ldots, \gamma_{K-1}$, which divides the real line into K intervals. When z_i lands in the k^{th} of these intervals, we observe $y_i = k$.
- Of course, neither the z_i's nor the cutpoints γ_k are known, so they must be inferred. This can be done using the class frequencies of the observed y_is though (many $y_i = k$ means the k^{th} bin is large).
- To model the influence of covariates $p\left(y_i = k|x_i\right)$, we suppose that $z_i = \beta^T x_i + \epsilon_i$. When ϵ_i is Gaussian, we recover ordinal probit regression, and when ϵ_i follows the logistic distribution[13] we recover ordinal logistic regression.
- An equivalent formulation of ordinal logistic regression models the "cumulative logits" as linear,

$$\log\left(\frac{p\left(y_i \le k\right)}{1-p\left(y_i \le k\right)}\right) = \alpha_k + \beta^T x_i. \tag{26}$$

  Here, the α_k's control the overall frequencies of the k classes, while β controls the influence of covariates on the response.
- Outside of the latent variable interpretation, it's also possible to understand β in terms of relative risks. In particular, when we increase the j^{th} coordinate by 1 unit, $x_i \to x_i + \delta_j$, the odds of class k relative to class $k-1$ gets multiplied by $\exp\left(\beta_j\right)$, for every pair of neighboring classes $k-1$ and k.

[13] This is like the Gaussian, except it has heavier (double-exponential-like) tails.

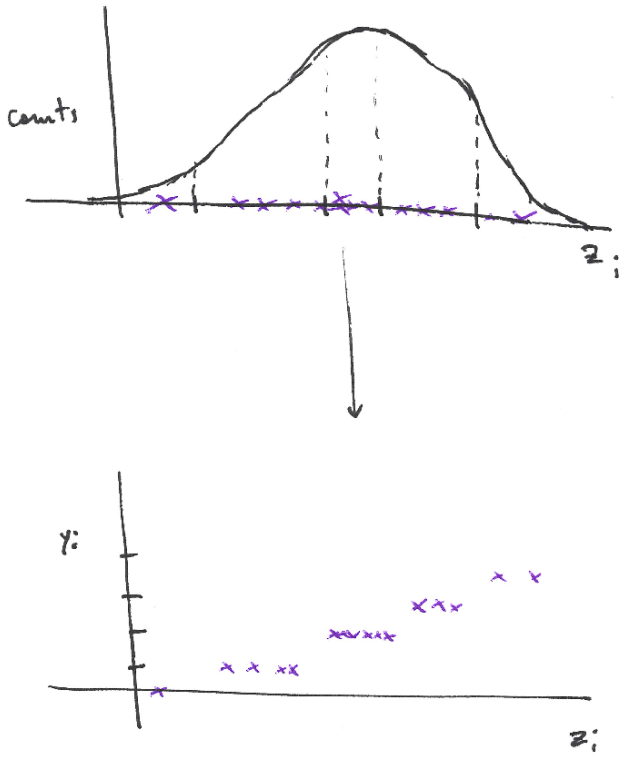

Figure 12: The association between the latent variable and the observed response in ordinal regression. The top panel is a histogram of the latent z_is. The dashed lines represent associated cutpoints, which determine which bin the observed response y_i belong to. The correspondence is mapped in the bottom panel, which plots the latent variable against the observed y_i – larger values of the latent variable map to larger values of y_i, and the width of bins in the top panel is related to the frequency of different observed classes. Latent variables give a natural approach for dealing with ordered categories.

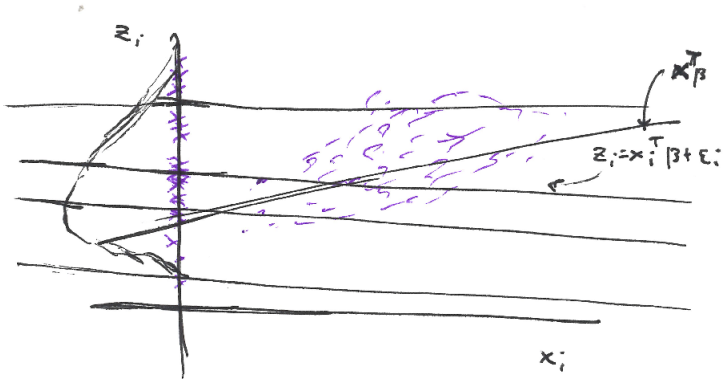

Figure 13: To incorporate the influence of features x_i in ordinal regression, a model is fit between the x_i and the latent z_i. The x-axis represents the value of this feature, and the y-axis represents the latent z_i. The histogram on the left is the same as the histogram in Figure 12. The (x_i, z_i) in the real data are in the purple scatterplot, and their projection onto the y-axis is marked by purple crosses. The bin in which the purple crosses appear (equivalently, which horizontal band the data point appears in) determines the observed class y_i. In this data, larger values of x_i are asociated with larger latent z_i, which then map to higher ordinal y_i.

5 Inference in linear models (and other more complex settings

It's worthwhile to draw a distinction between model estimation and statistical inference. Everything in section 4 falls under the purview of model estimation – we are building interpretable representations of our data by fitting different kinds of functions over spaces defined by the data. In contrast, the goal of inference is to quantify uncertainty in our analysis – how far off might our models be? The point is that critiquing the "insights" gained from a data analysis is just as scientifically relevant as the analysis itself. The testing and estimation procedures outlined in section 2 are a first step towards this kind of critical evaluation of data analysis, but those ideas become much more powerful when combined with GLMs.

5.1 (Generalized) Linear Models and ANOVA

One of the reasons linear models are very popular across many areas of science is that the machinery for doing inference with them is (1) relatively straightforwards and (2) applicable to a wide variety of hypotheses.

While we can fit linear models without making assumptions about the underlying data generating mechanism, these assumptions are crucial for doing valid inference.

- The high-level strategy of inference in linear models is to compare different types of data generating mechanisms, favoring those that are both parsimonious and fit the data reasonably well.

- As a consequence, we need to make assumptions about the data generating mechanism. Limiting our attention to linear models for now (GLMs are considered below), we assume the responses y_i are drawn according to

$$y_i = x_i^T \beta + \epsilon_i \tag{27}$$

  where x_i is considered fixed and ϵ_i are drawn i.i.d. from $\mathcal{N}\left(0, \sigma^2\right)$.

- Mathematically, the problem of comparing models can be approached by considering

$$\frac{RSS_{sub} - RSS_{full}}{RSS_{full}} \tag{28}$$

  where RSS_{full} and RSS_{sub} are residuals from fitting a full linear model and a submodel (the full model with some of the β_j's set to 0), respectively. The idea is that if the submodel doesn't lose much explanatory power (when compared ot the full model), then this quantity will be small, and we should favor the smaller model (Occam's razor). If it is large, however, then the improved fit of the full model is worth the cost in complexity.

- Suppose that the full and submodels have p_0 and p_1 parameters, respectively. Then, if the data are in fact generated by the submodel, it can be shown that

$$\frac{\frac{RSS_{sub} - RSS_{full}}{p_0 - p_1}}{\frac{RSS_{full}}{n - p_0}} \sim F_{p_0 - p_1, n - p_0}, \tag{29}$$

  since both the numerator and denominator are χ^2-distributed.

- This provides grounds for a formal hypothesis test: assume H_0 the data are generated from the submodel, but if the statistic in 29 is too large when compared to the reference F-distribution, reject it in favor of H_1 that the data are generated by the full model.

The two most common scenarios where this type of test is useful are,

- Testing $H_0 : \beta_j = 0$. When people talk about a variable in a linear model being significant, they mean that they have rejected this null hypothesis. The interpretation is that the submodel including all variables but this one does substantially worse than the full model that additionally includes β_j – note the connection to the interpretation of β_j as the amount that y changes when x_j is changed by a unit, conditional on all other variables (which make up the submodel). The results of this test are usually presented in terms of a t-statistic, but this follows directly from the fact that $t^2 \equiv F$ in this special case.

- Testing $H_0 : \{\beta_0 \in \mathbb{R}\} \cap \{\beta_1 = \cdots = \beta_p = 0\}$. This is the test of whether the model including the x_i's does better than the model which just predicts the average of the y_i's. Here, the submodel is the empty model.

Note that in an experimental setting (for example, where treatment and control were randomized), rejection of a submodel can be grounds for claiming causal relationships. This is because we have established that the response is nonconstant with respect to the terms not in the submodel, and all potentially confounding effects have been averaged out due to randomization.

A few variations on this standard linear model F and t-testing are worth being familiar with: ANOVA and testing for GLMs.

- ANOVA is actually just a name for a common special case of linear models, where there are only one or two predictors in x_i, all of which are categorical.

- One-way ANOVA refers to inference for models with one categorical variable,

$$y_i = \mu + \alpha_{l(i)}, \tag{30}$$

  where $l(i)$ gives the category level for sample i. The interpretation is that μ is a global average, while the $\alpha_1, \ldots, \alpha_L$ are the offsets for each category level[14].

- This model can be seen to be a special case of the linear model discussed above by writing the categorical levels in a $n \times L$-dimensional dummy coding matrix.

- Two-way ANOVA is exactly the same, except that now there are two categorical variables,

$$y_i = \mu + \alpha_{l_1^{(1)}(i)} + \alpha^{(2)}_{l_2(i)} \tag{31}$$

- For testing in GLMs, the key observation is we can replace the RSS with deviances, in equation 29, and the ratio is again F-distributed (though only approximately this time). This is not just some coincidence – they both arise from a ratio of loglikelihood ratio statistics, it's just that the linear model assumes a gaussian likelihood.

- From here, the derivation and interpretation of tests for different submodels (including those for individual β_j and for $\beta_1, \ldots, \beta_p$) are the same as those for ordinary linear models.

Finally, we note that this machinery for testing hypotheses directly provides a means of constructing confidence intervals for parameters: a confidence

[14] Note that this interpretation requires that we added a constraint the constraint $\sum_l \alpha_l = 0$ – the model is not identifiable without at least one constraint.

interval (or region) for a parameter (or a set of parameters) contains all the configurations of those parameters that we would not reject, if they were set as the values for the hypothesized submodel. It's better to recommend confidence intervals when possible, because they give an idea of practical in addition to statistical significance.

While understanding these fundamentals is important for using them successfully during a consulting session, the real difficulty is often in carefully formulating the client's hypothesis in terms of a linear model. This requires being very explicit about the data generating mechanism, as well as the pair / submodel pair of interest. However, this additional effort allows testing in settings that are much more complex than those reviewed in section 2, especially those that require controlling for sets of other variables.

- Testing the difference of mean in time series

5.2 Multiple testing

Multiple testing refers to the practice of testing many hypotheses simultaneously. The need to adapt classical methods is motivated by the fact that, if (say) a 100 tests were done all at a level of $\alpha = 0.05$, and the the data for each hypothesis is in fact null, you would still reject about 5 on average. This comes from the interpretation of p-values as the probability under the null hypothesis that you observe a statistic at least as extreme as the one in your data set.

5.2.1 Alternative error metrics

The issue is that in classical testing, errors are measured on a per-hypothesis level. In the multiple testing setting, two proposals that quantify the error rate over a collection of hypotheses are,

- Family-Wise Error Rate (FWER): This is defined as the probability that at least one of the rejections in the collection is a false positive. It can be overly conservative in settings with many hypothesis. That said, it is the expected standard of evidence in some research areas.
- False Discovery Rate (FDR): This is defined as the expected proportion of significant results that are false positives. Tests controlling this quantity don't need to be as conservative as those controlling FWER.

5.2.2 Procedures

The usual approach to resolving the multiple testing problem is (1) do classical tests for each individual hypothesis, and then (2) perform an adjustment of the p-values for each test, to ensure one of the error rates defined above is controlled.

Two adjustments are far more common than the rest,

- Bonferroni: Simply multiple each p-value by the number of tests N that were done (equivalently, divide each sigificance threshold by N). This procedure controls the FWER, by a union bound.

- Benjamini-Hochberg: Plot the p-values in increasing order. Look for the last time the p-values drop below the line $\frac{i\alpha}{N}$ (viewed as a function of i). Reject all hypotheses associated with the p-values up to this point. This procedure controls the FDR, assuming independence (or positive dependence) of test statistics among individual hypotheses.

A few other methods are worth noting,

- Simes procedure: This is actually always more powerful than Bonferroni, and applicable in the exact same setting, but people don't use it as often, probably because it's not as simple.
- Westfall-Young procedure: This is a permutation-based approach to FWER control.
- Knockoffs: It's possible to control FDR in high-dimensional linear models. This is still a relatively new method though, so may not be appropriate for day-to-day consulting (yet).

Here are some problems from previous quarters related to multiple testing,

- Effectiveness of venom vaccine
- Multiple comparisons between different groups
- Molecular & cellular physiology
- Pyschiatric diagnosis and human behavior

5.3 Causality

- Land use and forest cover

5.3.1 Propensity score matching

- The effect of nurse screening on hospital wait time

6 Regression variants

6.1 Random effects and hierarchical models

- Systematic missing test
- Trial comparison for walking and stopping

6.2 Curve-fitting

- Piecewise Linear Regression

6.2.1 Kernel-based

- Interrupted time analysis

6.2.2 Splines

6.3 Regularization

6.3.1 Ridge, Lasso, and Elastic Net

6.3.2 Structured regularization

6.4 Time series models

- UV exposure and birth weight
- Testing the difference of mean in time series
- Nutrition trends among rural vs. urban populations

6.4.1 ARMA models

6.4.2 Hidden Markov Models

- Change point in time course of animal behavior

6.4.3 State-space models

6.5 Spatiotemporal models

6.6 Survival analysis

6.6.1 Kaplan-Meier test

- Classification and survival analysis

7 Model selection

7.1 AIC / BIC

7.2 Stepwise selection

7.3 Lasso

- Culturally relevant depression scale

8 Unsupervised methods

Essentially, the goal of unsupervised methods is data compression, either to facilitate human interpretation or to improve the quality of downstream analysis. The compressed representation should still contain the signal in present the data, but the noise should be filtered away. Unlike regression and classification (which are *supervised*), no single variable is of central interest. From another perspective, unsupervised methods can be thought of as inferring latent discrete (clustering) or continuous (factor analysis) "labels" – if they were available, the problem would reduce to a supervised one.

8.1 Clustering

Clustering methods group similar samples with one another. The usual products of a clustering analysis are,

- Cluster assignments: Which samples belong to which clusters?
- Cluster characterization: How can we interpret the resulting clusters? This can be achieved by looking at cluster centroids (the within-cluster averages) or medoids (a sample from within the cluster that is representative of that cluster in some way).

Clustering techniques can roughly be divided up into those that are distance-based and those that are probabilistic. Common distance-based methods include,

- K-means: This builds K-clusters in an iterative way, attempting to minimize the sum of within-cluster distances (so that the clusters are compact). Since all directions are treated symmetrically, the resulting clusters tend to be spherical. Similarly, since clusters are treated symmetrically, they tend to all be approximately the same size.
- Hierarchical clustering: This builds clusters hierarchically. Specifically, it begins with each sample as its own singleton cluster. The pair that is found to be most similar is then merged together, and the procedure is continued until all clusters are joined into one large cluster. The nearness of a pair of cluster has to be defined somehow – common choices include "complete-link" merging (say that the distance between two clusters is the distance between the furthest away pair across clusters) or "average-link" (say the distance is the average pairwise distance between clusters). An advantage of this approach is that it provides a full tree relating samples – the more branch length a pair of samples shares, the more similar they are. In particular, this approach avoids the problem of choosing K, though clusters can be chosen by cutting the tree at some horizontal level. The main drawback of this approach is that it does not it does not scale as well as fixed-K methods.
- Spectral clustering: This

- Kernel k-means:

In these methods, decisions need to be made about (1) preprocessing and (2) the distance to use. Preprocessing will vary on a case-by-case basis. Generally, the goal of preprocessing is to ensure that the features included in clustering are meaningful, which means we might drop, transform (e.g., standardization), or derive features from (e.g., from counts to tf-idf scores) the original raw features. As far as distances are concerned, some useful ones to know are

- Euclidean distance: This is the standard length of a line between two points in a euclidean space: $d(x_i, x_j) = \sum_{k=1}^{p} (x_{ik} - x_{jk})^2$. It is a natural default, especially considering it's connection to squared-error loss / the exponent in the Gaussian. Note that, due to this connection to squared-error, it can be sensitive to outliers. For reference, the distance can be written as $\|x_i - x_j\|_2$.

- Weighted euclidean distance: If we want the coordinates to have different amounts of influence in the resulting clustering, we can reweight the coordinates in the euclidean distance according to

$$d(x_i, x_j) = \sum_{k=1}^{p} w_k (x_{ik} - x_{jk})^2, \tag{32}$$

  where w_k is the weight associated with coordinate k. This kind of reweighting can be useful when the relative weights for groups of columns should be made comparable to one another.

- Mahalanobis distance: This is defined as

$$d(x_i, x_j) = (x_i - x_j)^T \Sigma^{-1} (x_i - x_j) \tag{33}$$

  Note that ordinary and weighted euclidean distances are special cases, with $\Sigma^{-1} = I_p$ and $\operatorname{diag} w$, respectively. Geometrically, this distance reweights directions according to the contours defined by the ellipse with eigenvectors of Σ. Equivalently, this is the distance that emerges when using an ordinary euclidean distance on the "whitened" data set, which premultiplies X according to $\Sigma^{-\frac{1}{2}}$.

- Mannhattan / ℓ^1-distance: When applied to binary data, this counts the number of mismatches between coordinates.

- Cosine distance: This measures the size of the angle between a pair of vectors. This can be especially useful when the length of the vectors is irrelevant. For example, if a document is pasted to itself, the word count has doubled, but the relative occurrences of words has remained the same. Formally, this distance is defined as,

$$d(x_i, x_j) = 1 - \frac{x_i^T x_j}{\|x_i\|_2 \|x_j\|_2} = 1 - \cos\left(\theta_{x_i, x_j}\right), \tag{34}$$

  where θ_{x_i, x_j} represents the angle between x_i and x_j.

- Jaccard: This is a distance between pairs of length p binary sequences x_i and x_j defined as

$$d(x_i, x_j) = 1 - \frac{\sum_{k=1}^{p} \mathbb{I}(x_{ik} = x_{jk} = 1)}{p}, \tag{35}$$

  or one minus the fraction of coordinates that are 0/1. The motivation for this distance is that coordinates that are both 0 should not contribute to similarity between sequences, especially when they may be dominated by 0s. We apply this distance to the binarized version of the species counts.

- Dynamic time warping distance: A dynamic time warping distance is useful for time series that might not be quite aligned. The idea is to measure the distance between the time series after attempting to align them first (using dynamic programming).

- Mixtures of distances: Since a convex combination of distances is still a distance, new ones can be tailored to specific problems accordingly. For example, on data with mixed types, a distance can be defined to be a combination of a euclidean distance on the continuous types and a jaccard distance on the binary types.

In contrast, probabilistic clustering techniques assume latent cluster indicator z_i for each sample and define a likelihood model (which must itself be fit) assuming these indicators are known. Inference of these unknown z_i's provides the sample assignments, while the parameters fitted in the likelihood model can be used to characterize the clusters. Some of the most common probabilistic clustering models are,

- Gaussian mixture model: The generative model here supposes that there are K means $\mu_1, \ldots, \mu_K$. Each sample i is assigned to one of K categories (call it $z_i = k$), and then the observed sample is drawn from a gaussian with mean μ_k and covariance Σ (which doesn't depend on the class assignment). This model can be considered the probabilistic analog of K-means (K means actually emerges as the small-variance limit).

- Multinomial mixture model: This is identical to the gaussian mixture model, except the likelihood associated with the k^{th} group is Mult (n_i, p_k), where n_i is the count in the i^{th} row. This is useful for clustering count matrices.

- Latent Class Analysis:

- Hidden Markov Model: This is a temporal version of (any of) the earlier mixture models.The idea is that the underlying states z_i transition to one another according to a markov model, and the observed data are some emission (gaussian or multinomial, say) from the underlying states. The point is that a sample might be assigned to a centroid different from the closest one, if it would allow the state sequence to be one with high probability under the markov model.

Related to these probabilistic mixture models are mixed-membership models. They inhabit the space between discrete clustering continuous latent variable model methods – each data point is thought to be a mixture of a few underlying "types." These our outside of the scope of this cheatsheet, but see .... for details.

- Medwhat learning algorithm
- Unsupervised learning for classifying materials
- Clustering survey data
- CS 106A survey

8.2 Low-dimensional representations

8.2.1 Principle Components Analysis

- Relationship power scale
- Survey data underlying variables
- Analyzing survey data
- Teacher data and logistic regression
- Unsupervised learning for classifying materials

8.2.2 Factor analysis

- Culturally relevant depression scale

8.2.3 Distance based methods

8.3 Networks

- Molecular & cellular physiology
- Variational bounds on network structure

8.4 Mixture modeling

8.4.1 EM

9 Data preparation

9.1 Missing data

- Systematic missing test

9.2 Transformations

- Normalizing differences for geology data

9.3 Reshaping

10 Prediction

- Homeless to permanent housing

10.1 Feature extraction

- Classification based on syllable data
- Unsupervised learning for classifying materials

10.2 Nearest-neighbors

10.3 Tree-based methods

10.4 Kernel-based methods

10.5 Metrics

- Unsupervised learning for classifying materials

10.6 Bias-variance tradeoff

10.7 Cross-validation

- Classification and survival analysis
- UV exposure and birth weight
- Land investment and logistic regression

11 Visualization

The data visualization literature reflects influences from statistics, computer science, and psychology, and a range of principles have emerged to guide the design of visualizations that facilitate meaningful scientific interpretation. We'll review a few of those here, along with a few practical recommendations for how to brainstorm or critique data visualizations during a consulting session.

Some general principles developed by the data visualization community are

- Information density: Visualizations that maximize the "data-to-ink" ratio are much more informative than those that represent very little quantitative information using a large amount of space. For example, a histogram of a sample is more informative than just a mean + standard error. The flipside is tha reducing the amount of irrelevant "ink" (e.g., excessively dense grid lines) or white space can make it easier to inspect the true data patterns. The general philosophy is that an intricate but informative

visualization should be preferred to a simple-to-read but less-rich visualization.

- Small multiples: One way of increasing the information density within a figure is to create "small multiples." The idea is that if we want to make a certain kind of figure across many subgroups of a dataset, than it can be useful to make many small versions of that figure and place them side by side. For example, if we want to compare the time series of home prices across all the counties in a state, you can plot them simultaneously within small panels, and the essential shapes will be apparent (ordering them by some other characteristic would increase the information density even further). Practically, this can be achieved using `facet_grid` in ggplot2, for example.
- Focus + context: One difficulty in data visualization is that the same dataset can be visualized at several different scales. For example, in a network, we might want to study the neighborhoods of a few individuals as well as the overall community structure in that network. Focus + context refers to the practice of making it easy to compare views at different scales in the data. This can be done by using the same type of markers to refer to the same object across different figures or by using a smooth interactive transition.
- Linking: A generalization of focus + context is linking. Rather than comparing views of the same data at different scales, we can link views of different representations of the same data set. For example, if it is possible to view a data set as both a time series and a scatterplot, you could use some annotation to say that a particular point in the scatterplot refers to a full time series in the separate figure.
- Cognostics: The idea of cognostics is to make it easy to navigate a large collection of similar plots, much like in small multiples. In this setting however, it's impossible to view all the plots on one large panel. Instead, the idea is to compute a small set of summary statistics ("cognostics") for each plot, and then use that panel to sort / navigate across the whole collection.

12 Computation

12.1 Importance sampling

- Importance sampling

12.2 MCMC

12.3 Debugging peoples' code

- Reference for KS test

- Stata panel data