An overview of statistical methodology

Sunay Joshi

July 3, 2026

(Written to review STAT 9610, taught by Professor Eugene Katsevich. This is heavily based on his lecture notes.)

1 Frequentist

In frequentist statistics, we imagine that data is sampled from a data generating process. We don't know what the dgp is, so we postulate a statistical model, a family of probability distributions that hopefully contains the true dgp. Since "all models are wrong", we're faced at the outset with the fundamental tradeoff between robustness and efficiency. Should we confidently assert a very narrow parametric model and take advantage of this specificity, or instead use a broad nonparametric model at the expense of power? It depends.

For now, let's suppose that the model is correctly specified. Given n (x_i, y_i) pairs, the humble linear model postulates that $y \sim N(X\beta, \sigma^2 I)$ conditional on the design matrix X. That is, in the fixed-X setting, the responses are given by $y_i = x_{i*}^T \beta + \varepsilon_i$, where the errors ε_i are independent and identically distributed according to $N(0, \sigma^2)$, for some low-dimensional vector $\beta \in \mathbb{R}^p$ of regression coefficients. We typically assume X is full-rank; else, we lose identifiability of β, and the whole effort is pointless. Classical statistical models are always low-rank, low-dimensional; you can't fit more than n parameters with n datapoints. (You might protest that in modern life, we routinely fit overparametrized models, like DNNs. But regularization, be it explicit or implicit, reduces the effective dimensionality of such models. Otherwise they would clearly overfit the data.)

What do the regression coefficients mean? It is easiest to interpret the β_i's in the case that the predictors $x_{*1}, \ldots, x_{*p}$ (the columns of the design matrix) are orthogonal as vectors in $\mathbb{R}^n$. In this case, each predictor "explains" a different aspect of the response, and β_i is genuinely the effect of predictor i on the response. When the predictors are no orthogonal (i.e. we have "collinearity"), the previous sentence becomes: β_i is the effect of predictor i on the response, controlling for predictors $-i$. When there is high collinearity, i.e. some predictor i is nearly redundant, bad things happen: it becomes very hard to tell what β_i should be, so the variance of the estimate $\hat{\beta}_i$ blows up, and inference becomes conservative. More on all this below. We can adjust for the other predictors by a change of basis, but we don't do the details here.

We can estimate β, i.e. "fit the model", in many ways, the simplest being least-squares. The explicit formula for this estimator is $\hat{\beta} = (X^T X)^{-1} X^T y$, and its distribution is $N(\beta, \sigma^2 (X^T X)^{-1})$. The Gauss-Markov theorem states that under our rather stringent assumptions (!), the least-squares estimate is BLUE – the best linear unbiased estimate of β. (Linear as a function of y.) This formula can be simplified if the predictors are orthogonal: $\hat{\beta}_i \sim N(\beta_i, \sigma^2 (x_{*i}^T x_{*i})^{-1})$.

Much can be learned about least-squares estimation by thinking geometrically. Our model induces a model space $C(X)$ of mean vectors $\text{im}(X) = \{X\beta : \beta \in \mathbb{R}^p\}$, a p-dimensional subspace of $\mathbb{R}^n$. Our estimate $\hat{\mu} = X\hat{\beta}$ is obtained by orthogonally projecting the observed y onto $C(X)$. The R^2 of the regression, defined as the portion of variation in the response captured by the regression, is formally defined as $1 - RSS/TSS$, where RSS denotes residual sum of squares and TSS denotes total sum of squares. It's easy to see that R^2 is exactly the square of the cosine of the angle between y and $C(X)$, which in turn equals the squared empirical correlation between y and the fitted mean vector $\hat{\mu}$. A key point is that the fitted mean vector $\hat{\mu}$ only depends on X through the span $C(X)$; a change of basis for the predictors doesn't affect the fitted $\hat{\mu}$, though it does affect the interpretation of the components of $\hat{\beta}$. When we speak of "degrees of freedom", we're taking about the dimension of the model space $C(X)$, or later in the case of GLMs, its dimension as a submanifold of $\mathbb{R}^n$.

Now comes a central concept in statistics: inference about β. We would like to express the variability in our estimate $\hat{\beta}$ of β. Note that unlike in machine learning, where predictions are verified against observable outcomes, the parameter β is not observable. The quantity β never sees the light of day, and is just an artefact of our theorized model (unless of course, we are conducting a simulation). Inference is unique to statistics.

Inference can be done so long as we (at least approximately) know the distribution of our estimator. Inference comes in two forms: confidence intervals and hypothesis tests. A hypothesis test is a decision rule that either rejects or fails to reject the null, based on how surprising the data is. In the Neyman-Pearson framework, we are a bit more harsh on "Type I" errors, where we falsely reject the null. Under the null, the test must not incorrectly reject more than $100\alpha\%$ of the time. A confidence interval is a random interval that contains the true (deterministic) parameter β with a sufficiently high probability. Consider inference on β_j, the coefficient of predictor j. Controlling for the other predictors in the regression, is predictor j significantly associated to the response? Is the fitted value $\hat{\beta}_j$ larger than we would expect in the event that j been completely spurious? Inference always begins with a distributional statement of the form $\hat{\beta}_j \sim N(\beta_j, \sigma^2 [(X^T X)^{-1}]_{j,j})$. Then we construct the pivotal scalar quantity

$$T = \frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2 [(X^T X)^{-1}]_{j,j}}} \sim N(0, 1).$$

If σ^2 is known, then T is a z-statistic that can be used to test the null $H_0 : \beta_j = 0$. If σ^2 is not known, then T as written is infeasible. If we plug-in an unbiased estimate for σ^2, like

$$\hat{\sigma}^2 = \frac{1}{n-p} \sum_{i=1}^{n} (y_i - x_{i*}^T \beta)^2,$$

then the denominator becomes the standard error $\text{SE}(\hat{\beta}_j)$ of $\hat{\beta}_j$, and T becomes a t-statistic with distribution t_{n-p}, which can again be used to test $H_0 : \beta_j = 0$. (The Student t-distribution t_{n-p}

with $(n-p)$ degrees of freedom has slightly heavier tails than the standard normal, due to the fact that the extra randomness in the denominator adds variability to the statistic.) For a two-sided test, we reject the null if $|T| > t_{n-p}(1-\alpha/2)$. To construct a confidence interval, we note that

$$\left|\frac{\hat{\beta}_j - \beta_j}{\mathrm{SE}(\hat{\beta}_j)}\right| \leqslant t_{n-p}(1-\alpha/2)$$

with probability $(1-\alpha)$. Rearranging, we deduce that

$$\beta_j \in \mathrm{CI}(\beta_j) = [\hat{\beta}_j \pm t_{n-p}(1-\alpha/2) \cdot \mathrm{SE}(\hat{\beta}_j)],$$

our desired confidence interval. All this is for the highly special case of the normal linear; for more complicated models, like GLMs, there are in general three different inferential paradigms: Wald inference, likelihood ratio inference, and score inference. More on that when we discuss GLMs.

Now for a belated discussion of misspecification.

The most severe form of misspecification is model bias. Model bias is when our model for the conditional expectation of y given X is simply wrong; when the true regression function doesn't live in the function class we postulated. One natural fix is to pass to a richer function class; below, we discuss GLMs, but even GLMs can be limited (overdispersion necessitates Negative Binomial models). Omitted variable bias is a form of model bias. It is a frequent concern in causal inference. Suppose we are conducting a study of the effect of a certain medication for blood pressure. Suppose that in reality, the medication is a complete placebo, and has no effect on blood pressure at all. If we go ahead and regress the post-treatment blood pressure on an intercept and the indicator of treatment, the coefficient of the indicator should be near zero, right? Actually, this depends on how our control and treatment groups are constituted. Suppose that in the population, women have lower blood pressures on average than men. And suppose further that our treatment group had more women than men. Then although the drug has absolutely no effect on blood pressure, the treatment group would have a lower blood pressure than the control group post-treatment. (In this very simple scenario, we are not looking at the treatment effect, just the post-treatment outcomes.) If however we regress the outcome on the intercept, the indicator of treatment, and the indicator of gender, then the coefficient of treatment will indeed become zero. What went wrong? The issue was that gender and treatment assignment were correlated, due to bad study design. When we excluded gender from the regression, this forced the treatment variable to "do the heavy lifting" in "explaining" the observed response y. I imagine a dramatic scene between y and the predictors x_{*i}, in which y indignantly proclaims: "I demand an explanation from you! Intercept, what say you? Nothing? Well then, the culprit must be the treatment!" If the omitted variable was not correlated with the other predictors, this could instead appear in the error term. When designing a study, we need to be on the lookout for unsuspected correlations between treatment assignment and predictors.

There are less severe forms of misspecification. Nonnormality is what it sounds like; but for large sample sizes, this isn't much of an issue, thanks to the CLT. Heteroskedasticity is when the errors are independent and normal, but have different variances. It's quite common for larger responses to have larger noise. Huber-White standard errors are heteroskedasticity robust. Correlated errors are what the name suggests. In this case the effective sample size is diminished; an extreme case is when all the errors are equal, in which case the effective sample size is 1! If there is some sort of correlation decay, either in time or in space, then one can recover

from correlated errors. In time series data, one can use Newey-West standard errors. In data with group structure, Liang-Zeger standard errors do the trick.

Finally, there's the issue of outliers. We have assumed that our data is unblemished. But sometimes, people make mistakes, even at the level of data entry, and we would like our methods to be robust to such things. The first proposal might be to modify the loss function itself. The squared error loss, for all its mathematical convenience, is easily swayed by outliers. The minimum MSE estimate of a normal mean is the sample average, which gives equal influence $1/n$ to each datapoint – even abnormally large ones. On the other hand, the minimum absolute deviation estimator (based on the ℓ^1 loss) is the median, which is completely unaffected by outliers. Huber proposed a family of robust loss functions that transition from quadratic to linear at a certain distance δ from 0, and in the process developed the general theory of M and Z estimators (whose asymptotic "sandwich" variances bear his name). Aside from adjusting the loss function, one can try to single-out the outliers and remove these troublesome fellows from the data. However, this is a slippery slope; if you strategically remove enough points from your dataset, you can make the data say anything you like! But one can still try to define measures of "outlierness" and go from there; the general intuition is that outliers must be outside the bulk of x values and significantly above the regression line to have an effect. In high-dimensional data, even visually spotting outliers from low-dimensional projections can be a real challenge.

(A historical aside: in the 1970s, Princeton hosted a year on robustness, led by luminaries like Tukey (the founder of exploratory data analysis, among many other things) and Huber (whose name has already appeared multiple times). Those were the early days of statistical computing, which heralded the end of the golden era of mathematical statistics.)

(Another aside: we have been using the word "effect" liberally, but strictly speaking, we have only been studying association, not causation.)

The notion of a GLM begins with a generalization of the normal distribution. The normal location-scale family $\{N(\mu, \sigma^2) : \mu \in \mathbb{R}, \sigma^2 \geqslant 0\}$ is not an exponential family due to the σ^2, but is an instance of an exponential-dispersion model (EDM). In general, an EDM is a family of distributions parametrized by a natural parameter θ and dispersion ϕ of the form

$$f_{\theta,\phi}(y) = \exp\left(\frac{\theta^T y - \psi(y)}{\phi}\right) h(y).$$

In other words, it is a family of distributions on $\mathcal{Y}$ generated by exponential tilts of a base distribution $h(y)$. It's easy to see that the normal location-scale family fits the bill; as do many others, like the Poisson, Bernoulli, and Binomial proportion. In general, the mean and variance of an EDM are linked: the variance is given by $\phi_0 V(\mu)$ for some function V. For the normal EDM, V is the identity map, but for Poisson, $V(\mu) = \mu$, and for Bernoulli and Binomial proportions, $V(\mu) = \mu(1-\mu)$. In this sense, the normal EDM is but a tiny island of homoskedasticity in an ocean of heteroskedasticity; heteroskedasticity is the norm, not the exception.

With a more flexible family of distributions in hand, we can define a GLM as follows. We model y as distributed according to an EDM whose natural parameter is a linear function of the predictors, perhaps with an additional offset. As shown above, the natural parameter is related to the mean. Thus we are effectively modeling a certain link function of the mean as a linear combination of the predictors. We typically use the canonical link, in which case there is massive cancellation and $WM = 1$.

Geometrically, the model space is now the image of $\mathbb{R}^p$ under the inverse of the link function, a p-dimensional submanifold of $\mathbb{R}^n$. The fitted mean vector is no longer the orthogonal projection of y onto $C(X)$ with respect to the Euclidean inner product, but rather the projection with respect to a metric known as the total deviance.

We're again after β, but now its entries have different interpretations. Take for instance ungrouped logistic regression, where $y|x$ is modeled as $\text{Bin}(\mu)$, where $\text{logit}(\mu) = x^T\beta$. Here $\text{logit}(\mu) = \log \frac{\mu}{1-\mu}$ is the logit function, a homeomorphism from $(0,1)$ to $\mathbb{R}$. If $\beta_j = 3$, then a unit increase in the j-th predictor results in an increase of $e^3 \approx 20$ in odds of success.

As with linear models, we may perform maximum likelihood estimation to estimate $\hat{\beta}$. The score equation is now nonlinear, and can only be solved through iterative techniques. (We have the nonlinear equation $X^T WM(y-\hat{\mu}) = 0$, and we're after $\hat{\mu}$, but WM also depends on β; so fix WM, update $\hat{\mu}$, then fixing $\hat{\mu}$, update WM; and repeat, like a pair of pistons constantly alternating.) One technique is called iterative reweighted least-squares; here the weighting is due to the WM. (In EDMs, the presence of the w_i's puts different weights on the observations, not to mention the heteroskedasticity. Some observations are more noisy than others and thus contribute less to the estimate $\hat{\beta}$.)

Inference on the other hand is less straightforward. It's no longer immediate what the distribution of the least-squares estimate will be, given all the nonlinearity present. Here we rely on asymptotic theory; though we will discuss exact methods in finite sample, such as permutation tests. The notion of "asymptotic" will not be restricted to large sample theory, however; it will also refer to "small dispersion asymptotics", justified either by the CLT or by the saddlepoint approximation to EDMs.

As mentioned earlier, the three types of inference are Wald inference, likelihood ratio inference, and score inference. They coincide for the humble linear model but are quite different for arbitrary GLMs. In fact, we will see that Wald inference is far more problematic than its two counterparts. A nice memorization tool goes as follows: Wald inference makes use of the asymptotic distribution of the unconstrained estimate $\hat{\beta}$; likelihood ratio inference makes use of both the unconstrained and the constrained estimates; and score inference only relies on the constrained estimate. (I wonder if there is a universality result about these three forms of inference; aside from permutation tests, are all other likelihood-based techniques equivalent to one of these three?)

Wald inference makes use of the fact that the unconstrained MLE $\hat{\beta}$ is asymptotically normal with mean β and variance given by the inverse Fisher information $\phi_0(X^T WX)^{-1}$. This is a consequence of the general theory of the MLE, as discussed in STAT 9700. Now we mindlessly write down a test statistic as above. If we are testing a 1-dimensional parameter β_j, we write down either a z-statistic or a t-statistic, depending on whether we know ϕ_0. And if we are testing a multidimensional parameter β_S, we construct a quadratic form and compare to either the $\chi^2_{|S|}$ or $F_{|S|,n-p}$ distributions.

But Wald inference has a serious flaw, known as the Hauck-Donner effect. In the event that the true parameter lies close to the boundary of the parameter space – think $\pi \approx 0, 1$ for logistic regression – it turns out that Wald inference is conservative. This has to do with the fact that the standard error blows up faster than the parameter estimate. So even though Wald inference is typically the default procedure (even in R regression summaries), one must be careful. (As an

aside, Edgar told me that when the true parameter is near the boundary of the parameter space, the bootstrap also runs into trouble.)

Likelihood ratio based inference takes advantage of a different distributional fact. Suppose we are testing the null that $\beta_S = 0$. The null is $\beta_S = 0$, and the alternative is $\beta_S \neq 0$. Likelihood ratio inference fits the restricted MLE $\hat{\beta}^0$ under the null, as well as the unrestricted $\hat{\beta}$. By monotonicity, it is clear that $\hat{\beta}$ captures more of the variance in the response, but we ask whether it captures significantly more. The F-test compares the RSS of the restricted and the unrestricted fits, which under the null is distributed as $\chi^2_{|S|}$. In the LRT, the natural metric is no longer the squared error, but instead the unit deviance; thus we compare the total deviance $D(y, \hat{\mu}^0)$ of the restricted model to the total deviance $D(y, \hat{\mu})$ of the unrestricted model. Our LRT statistic is twice this difference, and can again be shown to be asymptotically $\chi^2_{|S|}$.

Score inference takes a critical look at the score function. I find the score function one of the most overlooked objects in statistics. I almost forget that it's a random function; in my eyes, it's always a means to an end, the MLE. In score inference, we make use of the fact that the score function evaluated at the restricted MLE $U(\hat{\beta}^0)$ is asymptotically normal. The actual test statistic is a nightmare to write down.

As a special case, we can consider goodness of fit (GOF) testing. In GOF, our null is that the data comes from a GLM, and the alternative is that the data comes from the saturated model whose mean vector is free to vary in all of $\mathbb{R}^n$. Thus null has p degrees of freedom, while the alternative has n. The basic idea is very simple: fit a GLM and see whether the residuals look "too large". LR-based GOF testing is based on the G^2 statistic given by

$$G^2 = \frac{1}{\phi_0} D(y, \hat{\mu}^0).$$

We reject if the residuals look too large, but size is now measured by the deviance, not squared error. Score-based GOF testing is based on the Pearson X^2 statistic given by

$$X^2 = \frac{1}{n-p} \sum_{i=1}^{n} \frac{w_i (y_i - \hat{\mu}_i^0)^2}{\phi_0 V(\hat{\mu}_i^0)}.$$

This is closer to the RSS, but weighted by the inverse of the variance. A mneumonic in the case of Poisson regression: $((O-E)/E)^2$. Both of these statistics are asymptotically χ^2_{n-p}, where $n-p$ is the difference in the number of degrees of freedom between the null and saturated models. Notice that we've written ϕ_0 as opposed to $\hat{\phi}_0$ in both the G^2 and X^2 statistics above. This isn't a typo: it's crucial to assume that we know the dispersion parameter ahead of time. If we used residuals to fit the dispersion parameter, and then re-used residuals to test GOF, we would never reject the null! Any piece of data should be used only once in a statistical procedure.

Before moving onto multiple testing, a note on permutation tests. Suppose we would like to test $H_0 : \beta_1 = 0$ in a simple linear regression $y = \beta_0 + \beta_1 x + \varepsilon$, where $x \in \{0, 1\}$ is categorical. Under the null, the two groups $x = 0$ and $x = 1$ are iid. In particular, the joint distribution of the datapoints is invariant under a permutation of the labels y. Consequently, for any test statistic T and any subset $A \subseteq S_n$ of permutations, one has that

$$S = \{T(X, \pi y) : \pi \in A \cup \{\mathrm{id}\}\}$$

is an exchangeable set of random variables. Under the null, with probability at least $(1-\alpha)$, the observed test statistic $T(X, y)$ lies below the empirical $(1-\alpha)$-quantile of the $|A|+1$ quantities in S; and so we reject if $T(X, y) > \mathbb{Q}_{1-\alpha}[S]$. This classical idea has made a comeback in recent years in the guise of conformal prediction, a distribution-free approach to predictive inference for blackbox models. One obtains calibration residuals that are exchangeable with the test residual, and by an argument analogous to the above, one can construct a predictive interval with finite-sample validity. It's awfully hard to say anything meaningful about the distribution of predictions coming out of a CNN, so if you want to gauge the uncertainty of your autonomous vehicle, you ought to use something like conformal prediction.

The final topic is multiple testing, also known as multiple comparisons. The subject was discussed as early as the 1950s by Tukey. The problem is very simply stated. Suppose you are testing m null hypotheses H_{0j} for $j \in [m]$. You compute p-values p_j for each of them. Your goal is to return a subset S of the hypotheses to reject.

If you naively used m level α tests, then even under global null, you would reject $m\alpha$ tests on average! A practical consequence of this is the reproducibility crisis in science. By testing a slew of hypothesis, scientists will eventually stumble upon a significant result by sheer chance, publish the finding in Nature, only to find that it doesn't replicate. What to do?

A breakthrough came in 1995 with the work of Benjamini and Hochberg. These developments came at a time when scientists were just beginning to work with truly big data; think genome wide association studies (GWASs) testing tens of thousands of genes for association with a particular disease. New methods were needed, and Benjamini and Hochberg delivered (accruing $100,000$ citations in the process, according to Google Scholar).

Let us fix concepts by defining the metric we wish to control: the false discovery rate (FDR). The quality of a given rejection set S is measured by the false discovery proportion (FDP), which quite intuitively is defined as the proportion of nulls among the hypotheses in S. (There is some ground truth as to which hypotheses are null and which are non-null; and we want to avoid rejecting nulls.) The FDR is simply the expected value of the random FDP, averaging over all sets of p-values we might sample.

Why is the FDR a natural metric? Why have we suddenly switched from Type I error control to this new and unfamiliar quantity? In the context of GWAS, the key point is that Type I error is not fatal, as the findings would inevitably be rigorously followed-up in a lab. But it is imperative that a large percentage be correct, to avoid wasting resources on false leads. In terms of the reproducibility crisis, if scientists all adopted the BH procedure at level $q = 0.10$, then among all published findings, 90% would be guaranteed to reproduce.

The simplest approach to FDR control is simply to adjust the level of each test from q to q/m. Then by a simple union bound, one can show that the resulting Bonferroni procedure controls FDR at level q. But in modern applications with a high multiplicity burden, q/m becomes tiny, and the resulting test loses all power to catch alternatives. Basically, most of the time, the set S will be empty. We need a less stringent criterion.

This is where the Benjamini-Hochberg (BH) procedure comes in. For any $q \in (0, 1)$, the BH procedure constructs S in such a way that the FDR is controlled at level q. It can be explained in one of two equivalent ways. The first is highly visual and from the original paper. Sort the p-values

from smallest to largest: $p_{(1)} \leqslant p_{(2)} \leqslant \ldots \leqslant p_{(m)}$. Since all p-values are to be treated equally, we ought to threshold the p-values at some point $\hat{t}$, and reject all those that fall below $\hat{t}$. But which $\hat{t}$ to choose? We will allow ourselves to be guided by the Bonferroni procedure, however flawed it may be. We plot j versus $p_{(j)}$, draw the Bonferroni line $j \mapsto jq/m$, locate the largest p-value $p_{(\hat{j})}$ that lies below the Bonferroni line, and set $\hat{t} = p_{(\hat{j})}$. It's that simple. The second exposition is more analytical. The FDR is an expectation of the random FDP, and the FDP depends on unobservable parameters (specifically, the null/non-null status of each hypothesis). How might we estimate the FDP? As above, we parametrize our rejection set $S = S(t)$ by our choice of threshold t, so that by definition,

$$\mathrm{FDP}(S(t)) = \frac{|S(t) \cap \mathcal{H}_0|}{|S(t)|},$$

where $\mathcal{H}_0$ denotes the subset of hypothesis that are truly null. Since we don't know the numerator, let's replace it with its expectation, which is at most $t|\mathcal{H}_0|$. Since we don't know the null proportion $|\mathcal{H}_0|/m$, we can further upper bound this by mt, arriving at the FDP (over)estimate

$$\widehat{\mathrm{FDP}}(S(t)) = \frac{mt}{|S(t)|},$$

and we define $\hat{t}$ as the largest t such that $\widehat{\mathrm{FDP}}(S(t)) \leqslant q$. Let us prove that the two formulations are equivalent. Note that the inequality $p_{(j)} \leqslant jq/m$ is equivalent to $mp_{(j)}/j \leqslant q$. When the threshold equals $p_{(j)}$, the set $S(t)$ is simply the set of j indices corresponding to the p-values that lie below the threshold. Thus the left-hand side is exactly $mt/|S(t)|$. The largest j such that $p_{(j)} \leqslant jq/m$ corresponds exactly to the largest t $(= p_{(j)})$ such that $mt/|S(t)| \leqslant q$. The actual proof of FDR control requires more work, and is only known under certain dependency conditions on the p-values.

One can also consider global tests, tests of global null $H_0 = \bigcap_{j=1}^m H_{0j}$. In that case, you want to control the Type I error, also known as the FWER. Your approach should depend on the dependency structure, and whether the signal (alternative) is concentrated (dense) or dispersed (sparse). In the dispersed case, you aggregate via combination tests, and you apply some clever transformation to the p-values. Bonferroni, Fisher combination, modern variants like Cauchy combination test (surprising mathematical statistics at play, stability of Cauchy), Vovk's harmonic mean test, assumptions on dependency structure, add...

To conclude, a few interesting mathematical tricks include Poissonization (relaxing the sample size to be a realization of a Poisson random variable, enabling us to use Poisson regression for independence testing in contingency tables); links between noncentral chi-squared distributions and central chi-squared distributions; and the backwards-martingale optional stopping proof of the validity of the BH procedure for independent p-values due to Storey et al.

2 Bayesian

(Not covered in STAT 9610.)

Quite different concepts

Bayesian inference

Bayesian linear regression (spike and slab)

Bayesian hypothesis testing (and paradoxes) (multiple testing?)

Conjugate priors

Heirarchical models

Importance sampling

MCMC versus VI

EM

Gaussian process models

Empirical Bayes

Prior and posterior checks

Simulating the data