Fundamentals of Statistics (18.6501x) review notes.

David G. Khachatrian

September 24, 2019

1 Preamble

This was made a good deal after having taken the course. It will likely not be exhaustive. It may also include some editorializing: bits of what I believe are relevant observations and/or information I have come across.

2 Conventions

We're using *denominator-layout notation* for matrix calculus. This convention suggests that ∇f is a column vector, i.e. for one-dimensional output f and a-dimensional input θ we'd have

$$\nabla_\theta f = \begin{bmatrix} \frac{df}{d\theta_1} \\ \frac{df}{d\theta_2} \\ \cdots \\ \frac{df}{d\theta_a} \end{bmatrix} \in \mathbb{R}^{a \times 1}$$

For multidimensional input, we stack the gradients horizontally:

$$\nabla_\theta g = \begin{bmatrix} | & \cdots & | \\ \nabla_\theta g_1(\theta) & \cdots & \nabla_\theta g_b(\theta) \\ | & \cdots & | \end{bmatrix} \in \mathbb{R}^{a \times b}$$

3 Introduction

("i.i.d" = "independent and identically distributed")

In probability theory, we are given the ground truth and try to understand the probability of certain observations. (e.g., $X \sim Geom(p)$, what is $Pr[X \geq 3]$?)

In *statistics*, we are given observations and try to understand the likelihood of a certain ground truth. (e.g., if we observe iid $X = (2, 3, 6, 3, 10)$ and we assume they come from a Geometric distribution, what p is most likely to have generated the data?)

4 Useful Probability Concepts

4.1 Moment Generating and Characteristic Functions

Consider e^X for a random variable X:

$$e^X = 1 + X + \frac{X^2}{2!} + \frac{X^3}{3!} + \cdots = f(X)$$

This looks like it could hold a lot of information about X's moments. We can't control what values X takes (it's a random variable after all), but we *can* insert another variable t and make a function of one r.v. and one regular variable:

$$e^{tX} = f(tX) = 1 + tX + \frac{(tX)^2}{2!} + \cdots$$

The *moment generating function* of a random variable X, if it exists, is described as

$$mgf_X(t) = E_X[e^{tX}] = 1 + tE[X] + \frac{t^2}{2!}E[X^2] + \cdots$$

The mgf has a useful property: if we take the derivative of the mgf k times and evaluate it at $t = 0$, we recover $E[X^k]$, i.e., $mgf_X^{(k)}(t = 0) = E[X^k]$.

Say X has PDF $f(x)$. Note that the mgf is essentially the Laplace Transform of $f(x)$. This sometimes diverges.

A larger set of functions will converge if you perform a *Fourier transform* on it (f simply need be L_1 integrable, which a probability must be essentially by definition since it must integrate to 1). In fact, this is used and called the *characteristic function* of a random variable X:

$$cf_X(t) = E_X[e^{itX}] = 1 + itE[X] + \frac{(it)^2}{2!}E[X^2] + \cdots$$

In this case, we have $cf_X^{(k)}(t = 0) = i^k E[X^k]$.

Note that *there is a one-to-one correspondence between probability distributions and their moment-generating functions (if they exist) and characteristic functions*. This means that $CF_X(t) \to CF_Y(t) \iff X \xrightarrow{(d)} Y$. (The pathological functions that might ruin such a correspondence cannot be probability distributions.)

4.2 Convergences

Almost Surely

$$T_n \xrightarrow[n\to\infty]{a.s.} T \iff Pr[\{\omega : T_n(\omega) \to_{n\to\infty} T(\omega)\}] = 1$$

In words, "No matter the event in the sample space Ω you can think of (if it has nonzero probability), it converges to the same value."

In probability

$$T_n \xrightarrow[n\to\infty]{\mathbb{P}} T \iff Pr[|T_n - T| \geq \epsilon] \to_{n\to\infty} 0 \; \forall \epsilon > 0$$

In words, "The probability of T_n being more than ϵ away from T goes to 0 as n increases."

In distribution *For all bounded and continuous functions f,*

$$T_n \xrightarrow[n\to\infty]{(d)} T \iff E[f(T_n)] \xrightarrow[n\to\infty]{E} [f(T)] \; \forall \text{bounded, continuous } f$$

In words, "For bounded f, the average of $f(T_n)$ converges to the average of $f(T)$ as n increases." (A simple example of why f must be bounded: It does not work with $f(X) = X$ if $X_n = \begin{cases} n, Pr = 1/n \\ 0 \text{ o.w.} \end{cases}$.)

4.3 Applications of Convergences

Law of Large Numbers (LLN) For iid X_i,

$$\bar{X}_n := \frac{1}{n}\sum_i X_i \xrightarrow[n\to\infty]{\mathbb{P},a.s.} E[X_1] = \mu$$

Central Limit Theorem (CLT) For iid X_i,

$$\sqrt{n}\frac{\bar{X}_n - \mu}{\sigma} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$$

Rough rule of thumb: Close enough for $n \geq 30$.

Hoeffding's Inequality For i.i.d $X_i \in [a,b](a.s.)$, *any* $n > 0$, we have an exponential tail bound on the sample average:

$$Pr[|\bar{X}_n - \mu| \geq \epsilon] \leq 2\exp\left\{-\frac{2n\epsilon^2}{(b-a)^2}\right\} \forall \epsilon > 0$$

Convergence of combinations of random variables. If $T_n \xrightarrow{a.s./P} T$, $U_n \xrightarrow{a.s./P} U$, then $T_n + U_n$, $T_n U_n$, and T_n/U_n converge almost surely/in probability to "what you'd expect".

Slutsky's Theorem If $T_n \xrightarrow{(d)} T$ and $U_n \xrightarrow{P} u$ for a constant u, then $T_n + U_n$, $T_n U_n$, and T_n/U_n converge *in distribution* (d) to "what you'd expect".

Continuous Mapping Theorem If f is continuous,

$$T_n \xrightarrow[n\to\infty]{a.s./P/(d)} T \implies f(T_n) \xrightarrow[n\to\infty]{a.s./P/(d)} f(T)$$

5 Foundations of Inference

The outcome of a statistical experiment provides a *sample* $X_1, \cdots, X_n \sim \mathbb{P}$ of n iid random variables. A *statistical model* is a pair

$$(E, (\mathbb{P}_\theta)_{\theta\in\Theta})$$

where E is the *sample space* (usually $\subset \mathbb{R}$), $(\mathbb{P}_\theta)_{\theta\in\Theta}$ is a *family of probability measures on* E, and Θ is a *parameter set*.

Note: it doesn't make sense to describe E using the parameters θ. For example, if describing $X_i \sim U[0,a]$ for unknown a, the model is $(\mathbb{R}_+, (U[0,a])_{a\geq 0})$.

A model is *well-specified* when the true probability distribution of X_i, $\mathbb{P}$, is contained by the family of probability measures of your model, $(\mathbb{P}_\theta)_{\theta\in\Theta})$. That is to say, $\exists\theta \in \Theta s.t. \mathbb{P} = \mathbb{P}$.

If $\Theta \subset \mathbb{R}^d, d \in \mathbb{N}^+$, the model is *parametric*. If Θ is infinite-dimensional, the model is *nonparametric*. If $\Theta = \Theta_p \times \Theta_{np}$, the model is *semiparametric*; we're interested in the finite dimensional Θ_p and the infinite-dimensional Θ_{np} is a *nuisance parameter*.

If a model is *identifiable*, then there is a injective mapping from θ to $\mathbb{P}_\theta$: (no two parameters can lead to the same probability distribution).

5.1 Estimator terminology

A *statistic* is any measurable function of a sample (essentially, if you can calculate it explicitly from the sample, it's a measurable function). An *estimator of* θ is a statistic whose expression doesn't depend on θ.

An estimator $\hat{\theta}_n$ of θ is *weakly/strongly consistent* if $\hat{\theta}_n \xrightarrow[n\to\infty]{P/a.s.} \theta$.

An estimator $\hat{\theta}_n$ of θ is *asymptotically Normal* if $\sqrt{n}(\hat{\theta}_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2)$. σ^2 is then called the *asymptotic variance of* $\hat{\theta}_n$.

$$bias(\hat{\theta}_n) = E[\hat{\theta}_n] - \theta$$

An estimator with zero bias is *unbiased*.

Since an estimator is a random variable, one can compute its variance $var(P) = E[(P - E[P]^2)] = E[P^2] - (E[P])^2$.

The *(quadratic) risk of an estimator* $\hat{\theta}_n$ is $R(\hat{\theta}_n) = E[(\hat{\theta}_n - \theta)^2] = Bias(\hat{\theta}_n)^2 + Variance(\hat{\theta}_n)$.

6 (Multivariate) Delta Method

Let $Z_n \in \mathbb{R}^a$ be a sequence of r.v. such that

$$\sqrt{n}(Z_n - \theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_a(0, \Sigma)$$

for $a \times a$ covariance matrix Σ.

If $g : \mathbb{R}^a \to \mathbb{R}^b$ (i.e., $g(\theta) = (g_1(\theta), \cdots, g_b(\theta)), \theta \in \mathbb{R}^a$) is *continuously differentiable at the point* θ, then we can say:

$$\sqrt{n}(g(Z_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_b\left(0, (\nabla_\theta g(\theta))^T \Sigma (\nabla_\theta g(\theta))\right)$$

7 Confidence intervals

A *confidence interval of level* $1 - \alpha$ *for* θ is a *random* interval (depending on sample X_i) $\mathcal{I}$ whose boundaries do not depend on θ and

$$\mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha \ \forall \theta \in \Theta$$

An *asymptotic* CI is the same as above, except asymptotically with increasing n, i.e.,

$$\lim_{n\to\infty} \mathbb{P}_\theta[\mathcal{I} \ni \theta] \geq 1 - \alpha \ \forall \theta \in \Theta$$

We can use the Normal approximation and some manipulation to write asymptotic intervals of the form

$$\hat{\theta}_n \in \left[\hat{\theta}_n - \frac{q_{\alpha/2}}{\sqrt{n}}\sqrt{Var(\theta)}, \hat{\theta}_n + \frac{q_{\alpha/2}}{\sqrt{n}}\sqrt{Var(\theta)}\right]$$

But these are not asymptotic *confidence* intervals because they depend explicitly on θ.
Some methods for fixing this:

1. **Most commonly used (if asymptotic CI):** Plug-in method. If using a consistent estimator, we have that $\hat{\theta}_n \xrightarrow{P/a.s.} \theta$. So $\frac{\hat{\theta}_n}{\theta} \to 1$. By Slutsky's Theorem, we find that we can simply "plug in" $\hat{\theta}$ where we see θ:

$$\hat{\theta}_n \in \left[\hat{\theta}_n - \frac{q_{\alpha/2}}{\sqrt{n}}\sqrt{Var(\hat{\theta}_n)}, \hat{\theta}_n + \frac{q_{\alpha/2}}{\sqrt{n}}\sqrt{Var(\hat{\theta}_n)}\right]$$

2. Take a conservative bound, if a maximum for the variance of the parameter is known. (For example, for Bernoulli models, $Var(p) \leq 1/4$.)

3. Solve explicitly for θ, if possible. The earlier interval implies that

$$\hat{\theta}_n - \frac{q_{\alpha/2}}{\sqrt{n}}\sqrt{Var(\theta)} \leq \theta \leq \hat{\theta}_n + \frac{q_{\alpha/2}}{\sqrt{n}}\sqrt{Var(\theta)}$$

   If you can solve for θ, you would have endpoints of a valid interval.

8 Hypothesis Testing

Let Θ_0 and Θ_1 be disjoint subsets of Θ. Then we can set up a hypothesis test:

$$\begin{cases} H_0 : \theta \in \Theta_0 \\ H_1 : \theta \in \Theta_1 \end{cases}$$

H_0 is the *null hypothesis*, H_1 is the *alternative hypothesis*. **These hypotheses do *not* play symmetric roles**: the data is only used to try to *reject* H_0.

We define a *test* $\psi \in \{0, 1\}$: a statistic which indicates whether we reject H_0. Often will take the form of $\psi = \mathbb{1}[T_n > C_\alpha]$ (or absolute-value around T_n) for some *test statistic* T_n (which presumably depends on the data, $T_n = f(X_1, \cdots, X_n)$) and value C_α. The *rejection region* is $R_\psi = \{x \in E^n : \psi(x) = 1\}$

When do we reject? Often, we control for the probability of accidentally rejecting H_0 when in reality $\theta \in \Theta_0$. This is called the *Type I error of a test* ψ:

$$\alpha_\psi : \Theta_0 \to [0, 1], \theta \mapsto P_\theta[\psi = 1]$$

We ensure this value is below a *(significance) level* α, i.e., $\alpha_\psi(\theta) \leq \alpha \ \forall \theta \in \Theta_0$.

We hope the probability of not rejecting H_0 even though in fact $\theta \in \Theta_1$ (*Type II error of a test* ψ) is low:

$$\beta_\psi : \Theta_1 \to [0, 1], \theta \mapsto P_\theta[\psi = 0]$$

We can describe the *power* of the test as $\pi_\psi = \inf \theta \in \Theta_1(1 - \beta_\psi(\theta))$. We hope to have high power.

The (asymptotic) *p-value* of a sample $x_1, \cdots, x_n$ for a given test ψ_α is the smallest (asymptotic) level α at which $\psi(x_1, \cdots, x_n) = 1$ (i.e., ψ rejects H_0).

9 Methods for Estimation

The most commonly used tool in statistics, the "statistical hammer" is estimating expectations using sample averages (knowing that they converge to the same value due to the Law of Large Numbers):

$$E[f(X)] \to \frac{1}{n}\sum_{i=1}^{n} f(X_i)$$

9.1 Differences between probability distributions

Total Variation Distance the total variation distance between two probability measures P_{θ_1} and P_{θ_2} is $TV(P_{\theta_1}, P_{\theta_2}) := \max_{A \subset E} \left| P_{\theta_1}(A) - P_{\theta_2}(A) \right|$. If E is discrete, we can show that $TV(P_{\theta_1}, P_{\theta_2}) = \frac{1}{2} \sum_{x \in E} \left| p_{\theta_1}(x) - p_{\theta_2}(x) \right|$ (similar expression if E is continuous).

A problem with the TVD is that if we compare a continuous probability distribution to a discrete one, the TVD is always 1. (For example, $X_i \sim Ber(p) \to N(0,1)$, but $TV(P_{\bar{X}_n}, P_{N(0,1)}) = 1$ no matter the size of n.)

Kullback-Leibler Divergence A more useful comparator between probability measures p and q is the *Kullback-Leiber (KL) divergence*:

$$KL(p,q) = \int_E p(x) \log\left(\frac{p(x)}{q(x)}\right) dx$$

(and analogously if E is discrete). Because the above is not symmetric and doesn't always satisfy the triangle inequality, it is not a distance but instead a divergence.

Fun fact: This is used in information theory, where it is called the *relative entropy* and describes the "cost" in Shannon entropy paid for encoding p using q instead. $H(p,q) = H(p,p) + KL(p,q)$.

Note that we can write $KL(p,q) = E_{p(x)}[\log\left(\frac{p(x)}{q(x)}\right)]$. Meaning, we can estimate it using our data!

Say the true parameter is θ. Since $KL(P_\theta, P_{\theta'})$ describes the closeness of the two distributions, it would make sense to set $\hat{\theta} = argmin_{\theta'} \hat{KL}(P_\theta, P_{\theta'})$. Some manipulation shows that in fact this is the same as maximizing the *likelihood* $L_n(x_1, \cdots, x_n, \theta')$ with respect to θ'.

Likelihood Given a statistical model $(E, (\mathbb{P}_\theta)_{\theta \in \Theta})$ and a sample of iid r.v. $X_1, \cdots, X_n$, the *likelihood of the model* is the map

$$L_n : E^n \times \Theta \to \mathbb{R}^{\geq 0},$$
$$(x_1, \cdots, x_n, \theta) \mapsto f_\theta(x_1, x_2, \cdots, x_n) \overset{i.i.d}{=} \prod_{i=1}^{n} f_\theta(x_i)$$

(PDF for continuous r.v.'s, PMF for discrete r.v.'s.)

To differentiate between *probability (densities)* and *likelihood*, consider that we are dealing with a function $f(\vec{x}, \vec{\theta})$. Loosely speaking, when we consider the observations fixed and let the model parameters vary, we are calculating the *likelihood (of the model given the data)*: $f(\vec{\theta}; \vec{x})$. When we consider the model parameters fixed and let the observations vary, we are calculating the *probability (of some observations, given the model)*: $f(\vec{x}; \vec{\theta})$.

9.2 Maximum Likelihood Estimation

The *maximum likelihood estimator* (MLE) of θ is defined as

$$\hat{\theta}_n^{MLE} = \underset{\theta \in \Theta}{\operatorname{argmax}}\, L(X_1, \cdots, X_n, \theta) = \underset{\theta \in \Theta}{\operatorname{argmax}} \log(L(X_1, \cdots, X_n, \theta))$$

The *Fisher information* can be written as

$$I_{n=1}(\theta) = Var_X\left(\nabla_\theta \left(\ln(L_{n=1}(\theta; X))\right)\right) = -E_X\left[\mathbf{H}_\theta \left(\ln(L_{n=1}(\theta; X))\right)\right]$$

(H is the Hessian, or "second-derivative matrix".)

In practice, you calculate the likelihood of the data $logL(x_i; \theta)$, and find the θ such that $\nabla_\theta L = 0$. This gives you the form of the MLE $\hat{\theta}^{MLE} = f(X_i, n)$.

9.2.1 Asymptotic Normality of MLE

If we have a number of conditions satisfied, we can guarantee asymptotic normality of the MLE. Let $\theta^* \in \Theta$. If

1. The model is identifiable.
2. For all $\theta \in \Theta$, the support of $\mathbb{P}_\theta$ doesn't depend on θ
3. θ^* is not on the boundary of θ
4. $\mathcal{I}(\theta)$ is (multiplicatively) invertible in a neighborhood of θ^*
5. (Some other technical conditions)

then:

$$\sqrt{n}\left(\hat{\theta}_n^{MLE} - \theta^*\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \mathcal{I}(\theta)^{-1})$$

9.3 Method of Moments

Usually, $E[X^j] = f(\theta)$. We also have that $\frac{1}{n}\sum_i X_i^j \to E[X^j]$ by the Law of Large Numbers. So, define a set of *sample moments* and note that:

$$\sqrt{n}\left(\begin{pmatrix}\hat{m}_1\\ \cdots \\ \hat{m}_d\end{pmatrix} - \begin{pmatrix}E[X]\\ \cdots \\ E[X^d]\end{pmatrix}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \Sigma)$$

with Σ being the covariance matrix between (population/sample) moments.

Consider the map M:

$$M : \Theta \to \mathbb{R}^d, \theta \mapsto M(\theta) = (m_1(\theta), \cdots, m_d(\theta))$$

If M is one-to-one, we can recover θ via $\theta = M^{-1}(m_1(\theta), \cdots, m_d(\theta))$.

So we define our *method of moments estimator* $\hat{\theta}_n^{MM} = M^{-1}(\hat{m}_1, \cdots, \hat{m}_d)$ if M^{-1} exists.

By considering $g = M^{-1}$, if M^{-1} is continuously differentiable at θ, we can use the Delta Method to get the asymptotic Normal distribution for $\sqrt{n}(\hat{\theta}^{MM} - \theta)$.

In general, the Method of Moments isn't as good as MLE, but if the MLE is intractable, the MM might be worthwhile (since the equations are polynomial).

9.4 M-estimation

M-estimation is a superset of MLE. Instead of maximizing $log(L_n) = \sum_i logL(X_i, \theta)$ w.r.t θ, we minimize $Q = \sum_i \rho(X_i; \mu)$ w.r.t μ for a function ρ you specify. The result is a minimizer $\hat{\mu}_n$, which is an estimator for the true minimizer μ^*. (If you assume a model with likelihood L, then setting $\rho = -log(L_1)$ recovers MLE exactly.)

9.4.1 Asymptotic Normality of m-estimator

Let $J(\mu) = \frac{\partial^2 Q}{\partial \mu \partial \mu^T}(\mu) \overset{\text{regularity conditions}}{=} E_{X_1}\left[\frac{\partial^2 \rho}{\partial \mu \partial \mu^T}(X_1; \mu)\right]$. Let $K(\mu) = Cov\left(\frac{\partial \rho}{\partial \mu}(X_1; \mu)\right)$. (For MLE, $J(\theta) = K(\theta) = \mathcal{I}(\theta)$.)

If:

1. μ^* is the only minimizer of Q.
2. $J(\mu)$ is invertible for *all* $\mu \in \mathcal{M}$
3. (Some other technical conditions)

then:

$$\sqrt{n}\,(\hat{\mu}_n - \mu^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(J(\mu^*)^{-1} K(\mu^*) J(\mu^*)^{-1}\right)$$

10 Hypothesis Testing Revisited.

10.1 The χ^2_d distribution

The χ^2 *distribution with d degrees of freedom* is the probability law governing the sum of iid $N(0,1)$ r.v.'s:

$$X = \sum_{i=1}^{d} Z_i^2,\ Z_i \overset{iid}{\sim} \mathcal{N}(0,1) \implies X \sim \chi^2_d$$

One can show that $\chi^2_d = Gamma(\alpha = \frac{d}{2}, \beta = \frac{1}{2})$. Another useful relationship is that

$$Z \sim \mathcal{N}_c(0, I_d) \implies \|Z\|_2^2 \sim \chi^2_d$$

10.2 Cochran's Theorem and the Student's T distribution

With some elbow grease, one can show *Cochran's Theorem*: for $X_1, \cdots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, we have

- For all n, $\sum_i X_i$ and $\sum_i (X_i - \bar{X}_n)^2$ are independent of one another.
- $\sum_i \left(\frac{X_i - \bar{X}_n}{\sigma}\right)^2 \sim \chi^2_{n-1}$

It's worth emphasizing that the X_i *must be Normally distributed*.

This can combine quite nicely with the *t-distribution with d degrees of freedom*, which describes the r.v.

$$\frac{Z}{\sqrt{V/d}}$$

where $Z \sim N(0,1)$, $V \sim \chi^2_d$, and $Z \perp V$ (they're independent of each other).

With quite a bit of finagling, we can show that

$$T_n = \frac{\bar{X}_n - \mu}{\sqrt{\frac{\sum_i (X_i - \bar{X}_n)^2}{n-1}}} \sim t_{n-1}$$

During a hypothesis test, we have an assumed value for the mean under the null μ_0, so that can replace the unknown μ. The rest is calculated from the data.

What if we had a two-sample test X_i and Y_j of n and m observations? If we assume the two samples are independent (and both Gaussian), we can sum their variances.

What would be the degrees of freedom for the t-distribution N for t_N? A conservative estimate is $N = min(n, m)$. Another method is the *Welch-Satterthwaite formula*:

$$N = \frac{(\hat{\sigma_a^2}/n + \hat{\sigma_a^2}/m)^2}{\frac{\hat{\sigma_a^4}}{n^2(n-1)} + \frac{\hat{\sigma_b^4}}{m^2(m-1)}}$$

(The value for N may be fractional.) This value is obtained by approximating the distribution of $\hat{\sigma^2}_x/n + \hat{\sigma^2}_y/m$ (with *unbiased* estimators for the variances) with the "closest" χ^2_N distribution. The derivation is a bit long but potentially worthwhile; see here for the derivation I posted on StackExchange.

10.3 Wald's test

Say we have a parametric model for a sample of X_i. Let's assume the MLE asymptotic Normality conditions for estimating a k-dimensional parameter θ are met. Then we had

$$\sqrt{n}\left(\hat{\theta}_n^{MLE} - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, \mathcal{I}(\theta)^{-1})$$

Let's make the right-side have covariance equal to I_k. We can do this by multiplying the left-hand side by $A = \mathcal{I}(\theta)^{1/2}$. By the Delta Method (taking the gradients..), this ends up multiplying the covariance by $A \cdot A^T$. Since the Fisher information is symmetric by construction, $A = A^T$ and the right side cancels out completely:

$$\sqrt{n}\mathcal{I}(\theta)^{1/2}\left(\hat{\theta}_n^{MLE} - \theta\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_k(0, I_k)$$

Now, let's perform the transformation $f(x) = x^T x = \|x\|_2^2$ on both sides (which we can do by the Continuous Mapping Theorem). We are now describing the sum of k iid standard Normal variables; that is to say, we're converging to χ^2_k:

$$T(n) = n\left(\hat{\theta}_n^{MLE} - \theta\right)^T \mathcal{I}(\theta)\left(\hat{\theta}_n^{MLE} - \theta\right) \xrightarrow[n\to\infty]{(d)} \chi^2_k$$

(We could also replace $\mathcal{I}(\theta) \to \mathcal{I}(\hat{\theta})$ because $\hat{\theta}$ is a consistent estimator of θ.)

Under a hypothesis testing framework where we assume we have a value $\theta = \theta_0$, we can calculate all the value on the left-hand side. Which means we have a test statistic! This is *Wald's test*. Since it has no control over the "direction of deviation" (very loosely, T_n calculates the "squared magnitude of deviation from θ_0"), it is best used for two-sided tests and not the best for one-sided tests (it'll likely be way too conservative).

10.4 Likelihood ratio test

Suppose instead you have a d-dimensional parameter θ and, rather than fixed all d parameters of θ for your null hypothesis, you only want to fix r of them:

$$\begin{cases} H_0 : \theta = (\theta_{fixed}, \theta_{free}), \theta_{free} \in \mathbb{R}^{d-r} \text{ is free to vary} \\ H_1 : \theta \neq (\theta_{fixed}, \theta_{free}) \end{cases}$$

As with essentially any hypothesis test with "leeway" for H_0, we tend to want to compare the "best" possible candidates for each class. So essentially, we want to compare a constrained MLE $\hat{\theta}_c$ (which represents

H_0) with the unconstrained MLE $\hat{\theta}$ (which represents H_1).

Consider the test statistic that compares the square of the ratio of log-likelihoods:

$$T_n = \log\left(\left(\frac{L_n(\hat{\theta})}{L_n(\hat{\theta}_c)}\right)^2\right)$$

By *Wilks's Theorem*, assuming H_0 is true and the MLE conditions for asymptotic Normality are met, then

$$T_n \xrightarrow[n\to\infty]{(d)} \chi^2_r$$

How did we determine the degrees of freedom? It's essentially

$$(\text{df for test}) = (\text{df for } H_1) - (\text{df for } H_0)$$

(Kind of makes sense, considering the name "degrees of freedom" and all.)

10.5 Implicit/Multiple Hypotheses

How would you deal with comparing multiple/implicit hypotheses where we don't fix parameters to specific values?

Describe your hypotheses in the following way:

$$\begin{cases} H_0 : g(\theta) = 0 \\ H_1 : g(\theta) \neq 0 \end{cases}$$

We can describe them as a function $g(\theta)$ and use the Delta Method. For example, $g(\theta) = (\theta_3, \theta_2 - \theta_1)$ tests whether $\theta_3 = 0$ and $\theta_1 = \theta_2$. You can then go on to perform e.g. Wald's test.

Sadly, none of these methods handle inequality constraint(s) in higher dimensions well. A seemingly promising paper on the subject can be found here.

10.6 Goodness of fit tests

A *goodness-of-fit test* sees whether a hypothesized *distribution* describes the data/observed samples well. This means we *don't* have a parametric model $(E, \mathbb{P}_{\theta,\theta\in\Theta})$ for the data: we need to describe both:

1. the *type* of distribution (e.g. Normal, Uniform, Poisson); this is a *nonparametric*, i.e., infinite-dimensional
2. the *parameters* for the distribution in question

Basically, we're testing for a point in function space, which is infinite-dimensional.

This sounds pretty difficult. In a way, it is. Most GoF tests don't have much power (i.e. they rarely reject the null, even when they should).

10.6.1 χ^2 test for discrete distributions.

A *categorical/Multinoulli distribution of degree K* is the extension of a Bernoulli distribution to K total outcomes. The distribution itself has $K-1$ degrees of freedom. Worth noting is that a properly designated categorical distribution can describe any PMF.

To compare whether discrete X_i came from a categorical distribution $C(p_1, \cdots, p_{k-1})$, you can use *the χ^2 goodness-of-fit test*, which is simply Wald's test applied to the categorical distribution:

$$T_n = \sum_{j=1}^{K} \frac{(\hat{p}_j - p_j)^2}{p_j} \xrightarrow[n\to\infty]{(d)} \chi^2_{K-1}$$

Here, $\hat{p}_j$ is the MLE for p_j and is simply $count(X_i = j)/n$.

Note the somewhat unexpected form of T_n: the summation includes the K'th component, which is entirely fixed by the first $K-1$ choices; and the denominator does not look like the "expected" variance $p(1-p)$. This all comes from applying Wald's test to this situation and performing a nontrivial amount of manipulation. (Hint: Define the categorical distribution using $p_1, \cdots, p_{K-1}$ and denote $p_K = 1 - \sum_{i=1}^{K-1} p_i$, i.e. p_K is not considered a parameter of the distribution. Now the parameters are all independent, and the Fisher information matrix will be invertible.)

10.7 Fun (and useful) facts about CDFs.

A possibly interesting fact is that for any X, $F_X(X) \sim U[0,1]$ (the CDF of a random variable is distributed uniformly). This actually makes sense when you think about it long enough. It's sort of a statement about quantiles: "If I randomly draw an X, it is equally likely to be the 25'th percentile as the 75'th percentile or 90'th percentile, etc."

If F_X is invertible, this can be used to simulate drawing from any distribution using the Uniform distribution:

1. Use a random number generator to get a sample u from $U[0,1]$.
2. The random sample from your target distribution is $F^{-1}(u)$.

Another possibly interesting fact is that the CDF of a r.v. X_i can be written as an expectation:

$$F_X(t) = Pr[X_i \le t] = E[\mathbb{1}[X_i \le t]]$$

So we can estimate it! Just replace $E[\mathbb{1}[X_i \le t]] \to \frac{1}{n}\sum_{i=1}^{n} \mathbb{1}[X_i \le t] = \hat{F}_n(t)$. This is known as the *empirical (or sample) CDF*. (We may drop the hat.)

What properties does our estimator have? By LLN, $\forall t \in \mathbb{R}, \hat{F}_n(t) \xrightarrow[n\to\infty]{a.s.} F(t)$. This is *pointwise convergence*, which sadly does not guarantee some nice properties we'd hope for. But we're in luck!

Glivenko-Cantelli Theorem (The Fundamental Theorem of Statistics) The estimator we have described above enjoys *uniform convergence to* $F(t)$, i.e.,

$$\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \xrightarrow[n\to\infty]{a.s.} 0$$

Donsker's Theorem We can keep going. By CLT, we could say $\sqrt{n}(\hat{F}_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, F(t)(1 - F(t)))$, but we can "do one better" (give a tighter bound) using *Donsker's Theorem*. If F is continuous, then

$$\sqrt{n}\sup_{t\in\mathbb{R}} \left|\hat{F}_n(t) - F(t)\right| \xrightarrow[n\to\infty]{(d)} \sup_{0\le t\le 1} |\mathbb{B}(t)|$$

where $\mathbb{B}(t)$ is a *Brownian bridge* on $[0,1]$.

10.7.1 The Kolmogorov-Smirnov test, Creating a Pivotal Distribution

So how do we piece it all together? If we have X_i drawn from unknown cdf F and we consider a *continuous* CDF F^*, we can have the hypothesis test:

$$\begin{cases} H_0 : F = F^* \\ H_1 : F \neq F^* \end{cases}$$

For the asymptotic case, we can simply appeal to Donsker's theorem and use the

$$T_n = \sqrt{n} \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - F^*(t) \right|$$

and use tabulated results for the Brownian bridge to figure out what T_n is. To calculate T_n, compare the points of discontinuity of $\hat{F}_n(t)$ with the corresponding points in F^*, and take the max.

What about for finite n? We can use the fact that $F_X(X) \sim U[0,1]$. The process is as follows:
Repeat the following many times:

1. Generate n iid samples $Y_i \sim U[0,1]$.
2. Sort $Y_i \to Y_{(i)}$.
3. Describe $\hat{F}_n(\frac{i}{n}) = Y_{(i)}$.
4. Calculate $T_n = \sqrt{n} \times \max_{i \in \{1,\cdots,n\}} (\hat{F}_n(\frac{i}{n}) - \frac{i}{n})$
5. Store resulting value (in, e.g., list L).

After generating this *pivotal* distribution (so-called because it doesn't depend on parameters we don't know), compute T_n for the sample you actually observed:

1. Calculate $Y_i \leftarrow F^*(X_i)$.
2. Sort $Y_i \to Y_{(i)}$.
3. Describe $\hat{F}_n(\frac{i}{n}) = Y_{(i)}$.
4. Calculate $T_n = \sqrt{n} \times \max_{i \in \{1,\cdots,n\}} (\hat{F}_n(\frac{i}{n}) - \frac{i}{n})$
5. Compare T_n to values in list L generated above. The quantile of T_n determines its p-value: $p = 1 - \text{quantile}$.

10.7.2 Kolmogorov-Lilliefors test

What if we just want to know if our data come from a Normal distribution? If we use a test statistic

$$T_n = \sqrt{n} \sup_{t \in \mathbb{R}} \left| \hat{F}_n(t) - \Phi(t; \hat{\mu}, \hat{\sigma}^2) \right|$$

we need to use a different test than the KS test, because we have partially fitted our proposed distribution to the data. The appropriate test is the *Kolmogorov-Lilliefors test*. It is more stringent because of the partial fitting: $F_{KL}(t) > F_{KS}(t)\ \forall t$, which means the same-valued test-statistic will have a smaller p-value for KL compared to KS.

10.7.3 Quantile-Quantile Plots

Even though we talked about all this, very often we simply visually inspect proximity to a desired distribution using *quantile-quantile plots*. If we want to see whether X_i may have F as its governing distribution, we plot $X_{(i)} = F_n^{-1}(\frac{i}{n})$ on the Y-axis, and $F^{-1}(\frac{i}{n})$ on the X-axis:

$$(x_i, y_i) = (F^{-1}(\frac{i}{n}), F_n^{-1}(\frac{i}{n}))$$

If the resulting line is near $y = x$, then it's probably OK.

We can consider relative weights of tails using a Q-Q plot. Consider the top-right quadrant of the Q-Q plot. If points are to the right of $y = x$, then $y < x$. The target distribution's q'th quantile is larger than the sample's q'th quantile. This suggests that the target distribution has thicker tails than the sample distribution. One can make similar inductions for other patterns.

11 Linear Regression

(We have discussed Bayesian statistics in depth in the Probability review sheet and so do not repeat ourselves here.)

Consider a sample of (X_i, Y_i) pairs for a random variable/vector X and another variable Y. A common goal is to understand the joint distribution $\mathbb{P}$ of (X, Y); or (perhaps more often) the conditional distribution of Y given X.

We could model $\mathbb{P}$ *entirely* by trying to find the joint PDF $f_{X,Y}(x, y)$ and the conditional PDF of Y given X, $f_{Y|X}(y \mid x) = \frac{f_{X,Y}(x,y)}{\int_{y \in Y} f_{X,Y}(x,y)dy}$.

We often model $\mathbb{P}$ *partially*. Most often[1], we attempt to describe the conditional expectation of Y given X, called the *regression function* $\mu(x)$:

$$x \mapsto \mu(x) := E_Y[Y \mid X = x] = \int_{y \in Y} y f_{Y|X}(y \mid x)dy$$

The space $\mathcal{F}$ of possible regression functions f is nonparametric (infinite-dimensional), so we should probably limit ourselves in some way. If we assume that μ is an affine function:

$$\mu(x) = a + bx,\ x \in \mathbb{R}^1 \text{ (univariate regression)}$$
$$\mu(x) = x^T\beta,\ x \in \mathbb{R}^{d+1},\ d \in \{1, 2, \cdots\} \text{ (multivariate regression)}$$

we are talking about *linear regression*. We first discuss univariate regression, then extend to multivariate regression.

11.1 Univariate regression

One can perform theoretical regression, i.e. find the optimal parameters a^*, b^* such that

$$(a^*, b^*) = \underset{(a,b) \in \mathbb{R}^2}{\text{argmin}}\ E\left[(Y - (a + bX))^2\right]$$

You can solve the minimization problem (set gradient to zero, ...) to find that

$$b^* = \frac{Cov(X, Y)}{Var(X)}, \quad a^* = E[Y] - b^*E[X]$$

[1] In other cases, we may be more interested in the conditional median/α-quantile, conditional variance, etc.

The resulting line, which minimizes the *sum of squared residuals (SSR)* (technically the average squared residual above, but the difference is a factor of n), doesn't fit through an observed sample perfectly. We can call the residuals *noise* $\epsilon = Y - (a^* + b^*X)$. So we can write

$$Y = a^* + b^*X + \epsilon$$

where, by construction,

1. $E[\epsilon] = 0$
2. $Cov(X, \epsilon) = 0$

Now, in practice, we don't know the true ("population") SSR, expectations, variances, and covariances for X and Y ($E[X], Cov(X, Y), \cdots$). But we have samples X_i and Y_i so we can estimate all of them! That is to say, we now solve:

$$\begin{aligned}(\hat{a}, \hat{b}) =& \underset{(a,b)\in\mathbb{R}^2}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}\left[(Y_i - (a + bX_i))^2\right] \\ =& \underset{(a,b)\in\mathbb{R}^2}{\text{argmin}} \sum_{i=1}^{n}\left[(Y_i - (a + bX_i))^2\right]\end{aligned}$$

and we get our estimators $\hat{a}$ for a^* and $\hat{b}$ for b^* which are esseentially plug-in estimators of the theoretically optimal values described above ($Cov(X, Y) \to \bar{XY} - \bar{X}\bar{Y}, \cdots$). (We can drop the $1/n$ factor because the argmin does not change when we multiply/divide by a constant positive factor. Now we're truly minimizing the observed SSR.) We end up with an estimate for the optimal linear model:

$$Y_i = \hat{a} + \hat{b}X_i + \epsilon_i$$

Remember that ϵ is also a random variable! We can ask questions like "Does it look Gaussian?" (which would use a variant of the KL test). We could also simply make an assumption about its form: most commonly, that it's Gaussian $N(0, \sigma^2)$. In which case, we have observations of ϵ, so we can estimate its variance σ^2, etc.!

11.2 Multivariate Regression

In the multivariate case, $X' \in \mathbb{R}^d, d \geq 1$ is a random vector. We augment X' by prepending a 1 to the vector (to simplify notation). So $[1|X_i'] = X_i \in \mathbb{R}^{d+1}$. Let's define $p := d + 1, p \geq 2$. By previous notation, X is a column vector.

Now define a *design matrix* $\mathbb{X} \in \mathbb{R}^{\times}$ comprised of each observations X_i stacked vertically on top of each other (i.e., each row corresponds to one observation). Let $\beta \in \mathbb{R}^p$ be a column vector, $Y \in \mathbb{R}^n$ be the concatenation of observations on the target variables (aligned with the appropriate row in the design matrix), and $\epsilon \in \mathbb{R}^n$ be the residuals per observation pair. Now we have the relationship:

$$Y_i = X_i^T\beta + \epsilon_i, \quad Y = \mathbb{X}\beta + \epsilon$$

And we determine β by minimizing the SSR:

$$\begin{aligned}\hat{\beta} =& \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \frac{1}{n}\sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 \\ =& \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \sum_{i=1}^{n}(Y_i - X_i^T\beta)^2 \\ =& \underset{\beta\in\mathbb{R}^p}{\text{argmin}} \|Y - \mathbb{X}\beta\|_2^2\end{aligned}$$

The resulting estimator is the *least squares estimator of* β, $\hat{\beta}^{MLE}$. An analytic computation (take gradient wrt β, set equal to zero, solve...) shows that

$$\hat{\beta}^{LSE} = (\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T Y$$

A geometric interpretation: $\mathbb{X}\hat{\beta}$ is the orthogonal projection of Y onto the subspace spanned by the columns of $\mathbb{X}$:

$$\mathbb{X}\hat{\beta} = PY, \quad P = \mathbb{X}(\mathbb{X}^T\mathbb{X})^{-1}\mathbb{X}^T$$

11.2.1 Gauss-Markov Assumptions and subsequent properties

Let's assume the following are true:

1. The design matrix is deterministic and $rank(\mathbb{X}) = p$. (This suggests you control and decide in advance which samples X_i you get to see. In reality, this is not often the case, but you can instead say "conditioned on $\mathbb{X} = X_{observed}$, ...")
2. The model is *homoskedastic*, i.e. ϵ_i are iid
3. The noise vector $\epsilon \sim N_n(0, \sigma^2 I_n)$ for some (known or unknown) $\sigma^2 > 0$

(These are the *Gauss-Markov assumptions*.) Then:

1. $Y \sim N_n(\mathbb{X}\beta^*, \sigma^2 I_n)$
2. $\hat{\beta} \sim N_p(\beta^*, \sigma^2(\mathbb{X}^T\mathbb{X})^{-1})$
3. Quadratic risk of $\hat{\beta}$: $E[\|\hat{\beta} - \beta\|_2^2] = \sigma^2 trace((\mathbb{X}^T\mathbb{X})^{-1})$
4. Prediction error: $E[\|Y - \mathbb{X}\hat{\beta}\|_2^2] = \sigma^2(n-p)$
5. An unbiased estimator for σ^2 is $\hat{\sigma}^2 = \frac{\|Y - \mathbb{X}\hat{\beta}\|_2^2}{n-p} = \frac{\sum_i \hat{\epsilon_i^2}}{n-p}$.
6. $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$
7. $\hat{\beta} \perp \hat{\sigma}^2$.

This laundry list of results come from (1) algebra/manipulation/reasoning, (2) previous results (e.g. Cochran's Theorem).

One thing that may help elucidate the others (though the explanation may need to be cleaned up): the chi-squared distribution has $n-p$ degrees of freedom. We end up having p degrees of freedom when choosing the best estimator $\hat{\beta}$, but the original *sample space* in $\mathbb{R}^n$ (which we were in prior to projecting down to β-space $\in \mathbb{R}^p$) has n dimensions. So ϵ^2 only gets to wiggle around freely along $n-p$ dimensions; the rest are "pegged down" because we chose $\hat{\beta}$ in order to minimize the SSR.

11.2.2 (Multiple/Implicit) Hypothesis testing on parameters.

How would you test hypotheses on these parameters? With the Gauss-Markov assumptions, we have Normally distributed $\hat{\beta}$ and chi-squared distributed $\hat{\sigma}^2$. So we can perform t-tests! What if you're doing an implicit test? Wald's test and the Delta Method! For example,

$$\Big\{ H_0 : \beta_1 = \beta_3^2 H_1 : \beta_1 \neq \beta_3^2$$

You can define $g(\beta) = \beta_3^2 - \beta_1$, take its gradient, apply Delta Method, etc.

What about multiple hypotheses at once, e.g. $\forall i, \beta_i = 0$? You'll need to be more careful to ensure you control the signficance level properly. A simple but very conservative method to achieve *familywise error rate (FWER)* α is to test each of the K hypotheses at level α/K. This is called the *Bonferroni correction*.

A method that controls the FWER controls the probability that *any* of the hypotheses are false positives. You'll probably want a less excessively conservative method, which instead constrols for the *false discovery rate (FDR)*, which describes the fraction of positives that are false positives. One example of a method that controls the FDR is the Benjamini-Hochberg method, described here.

11.2.3 Other types of estimators.

We have been minimizing an objective function J to get our estimator $\hat{\beta}$:

$$\underset{\beta}{\text{argmin}}\ J(X, Y, \beta)$$

When we chose J to be $\|Y - \mathbb{X}\beta\|_2^2 = SSR(X, Y, \beta)$, we'd get the least-squares estimator $\hat{\beta}^{LSE}$. Under the Gauss-Markov assumptions, this is $\hat{\beta}^{MLE}$.

But what if we want to minimize a *different* function? Well, that's basically M-estimation! (If the conditions are met, we can appeal to some of those results as well!) We'll generally write our objective function as

$$J(X, Y, \beta, \lambda) = L(X, Y, \beta) + \lambda R(\beta)$$

where L is a *loss function* that describes how well/poorly our model fits the data, R is a *regularization function* that penalizes certain values of β more heavily than others, and λ is the *regularization factor* that determines the relative importance of L and R; larger λ places more importance on R compared to L.

(Note: Technically, what we're "officially" doing is minimizing L subject to the constraints described by R, and then we use the method of Lagrange multipliers to recast the optimization function to the "soft" form described above. This works fine; there would be a correspondence between the expression for λ (frequentist viewpoint) and the prior placed on β (Bayesian viewpoint). See more below.)

Some examples/names of estimators with special names. One should consider the estimators' *risk*, *variance*, *bias*, etc and the use-case at hand when deciding which is best:

1. The *ordinary least squares (OLS)* estimator is the same as the "regular" least-squares estimator; $R = 0$ and $L = \|Y - X\beta\|_2^2$.

2. The *ridge regression* estimator has $R(\beta) = \|\beta\|_2^2$. This lowers the "energy" of the estimator, and might provide enough curvature to J to create a unique minimizer. (Otherwise, collinearity among features will "break" the estimator and/or greatly increase variance.)

3. The *least absolute shrinkage and selection operator (LASSO)* estimator has $R(\beta) = \|\beta\|_1^1$. This is a soft form of lowering the "dimensionality" of the estimator.

4. The *principal component regression (PCR)* estimator instead considers the following: Consider the eigendecomposition $X^T X = PDP^T$. Now take the first k columns of P and , called P_k ($p \times k$). Now describe $\hat{\beta} = P_k \hat{\gamma}$, where $\hat{\gamma}$ is found by minimizing $L = \|Y - XP_k\gamma\|_2^2$. This is essentially fitting β onto the k-dimensional subspace that is "closest" to the p-dimensional space spanned by X. This is another way to ensure multicollinearity or uninformative features do not "break" an OLS estimator (having huge variance, etc).

11.2.4 Connection with Bayesian inference.

There is in fact a really strong connection with Bayesian inference at play here, especially when we are assuming a model for β. To make the comparison clearer, let's assume we want to *maximize* J instead of minimize it (by doing $J' = -J$, we get the exact same answers, so our problem hasn't actually changed — we just have fewer negative signs to think about).

Consider a Bayesian inference setup. You have a random variable Θ you want to estimate. You have a prior belief $\pi(\theta)$, then you observe $X_i, Y_i \sim \mathbb{P}_\Theta$ and use its likelihood function $L_n(X, Y \mid \theta)$ to update to your posterior distribution:

$$\pi(\theta \mid X_i, Y_i) = const \times L_n(X, Y \mid \theta)\pi(\theta) \propto L_n(X \mid \theta)\pi(\theta)$$

Normally, after this calculation, you'd do whatever you want with your posterior distribution $\pi(\theta \mid X_i, Y_i)$. If all you were looking for was the *maximum a posteriori (MAP)* estimator for Θ, $\hat{\Theta}^{MAP}$, you could simply maximize

$$\max_{\theta \in \Theta} \pi(\theta \mid X_i, Y_i) = L_n(X \mid \theta)\pi(\theta)$$

Now simply take the log. (This gives the same optimizing argument.) Now you're maximizing:

$$\max_{\theta \in \Theta} \log(\pi(\theta \mid X_i, Y_i)) = \log(L_n(X \mid \theta)) + \log(\pi(\theta))$$

Now compare! The log-likelihood plays the role of the loss function: $\log(L_n(X, Y, \theta)) \leftrightarrow L(X, Y, \beta)$, and the prior belief plays the role of the regularization function: $\pi(\theta) \leftrightarrow R(\beta)$, with λ balancing the relative importance of each. So when you place a regularization term in an objective function, you're implicitly indicating a prior belief on the parameter space! (Which I'd say counts as "pretty darn neat".)

12 Generalized Linear Models (GLMs)

In the previous section on linear models, we made two main assumptions to get the results we obtained:

1. The target variable given the data was Gaussian: $Y \mid X = x \sim N(\mu(x), \sigma^2)$
2. The regression function was linear: $\mu(x) = x^T\beta$.

But there are cases where these assumptions don't make much sense. For example, if the target variable $Y \mid X = x \sim Ber(p)$. We have $p = E[Y \mid X = x] = \mu(x) \in (0, 1)$, so clearly $\mu(x) \neq x^T\beta$ (which has range $\mathbb{R}$).

Changes in x should affect the mean of Y, and we should be able to describe the change additively in some way. But how? Probably not through an additive change to p directly, because than we can have $p < 0$ or $p > 1$. It might make more sense to say it increases the *log-odds* of something occurring, so that it's additive to $\log\left(\frac{p}{1-p}\right)$. How do we fit this notion to our linear model framework above?

This is where the *link function* g comes in. In a generalized linear model (GLM):

1. $Y \mid X = x \sim$ a member of exponential family
2. We choose a monotone increasing and differentiable function $g : Range(Y) \to \mathbb{R}$ so that $g(\mu(x)) = x^T\beta$. g is the *link function*.

We choose the distribution based on information we have about the problem at hand, as long as it's a member of an exponential family (defined below). From there, we choose a link function g that makes $g(\mu(x))$ linear in β (though their coefficients $k(x)$ need not be linear in x).

Why restrict our choice of distribution to members of an exponential family? This allows us to claim that *the asymptotic Normality of the MLE also applies* to such models (if technical conditions are met), allowing us to perform hypothesis tests on our parameters, etc.

12.1 Exponential Families

OK, so what's an exponential family?

A family of distributions $\{\mathbb{P}_\theta : \theta \in \Theta\}, \Theta \subset \mathbb{R}^k$ is a *k-parameter exponential family on* $\mathbb{R}^q$ if there exists the following real-valued functions:

1. *natural parameters of* θ: $\eta_1(\theta), \cdots, \eta_k(\theta)$
2. *sufficient statistics of the target variable's observations* Y: $T_1(z), \cdots T_k(z)$
3. bias functions of θ and z: $B(\theta)$, $h(z)$

Which allow us to write

$$f_\theta(z) = \exp\left[-B(\theta) + \sum_{i=1}^{k} \eta_i(\theta) T_i(z)\right] h(z)$$

Basically, every function of the parameters and the observations can be decomposed into a product: $K(\theta, z) = f(\theta)g(z)$.

What does any of this have to do with our GLM?

We observe pairs (X_i, Y_i). θ is the parameter(s) of our the distribution modeling $Z_i = Y_i \mid X_i$. We need to be able to write the components of θ in terms of $E[Z_i] = E[Y_i \mid X = x_i]$.

Next, we have chosen a link function for our regression function so that $g(\mu(x_i)) = g(E[Y \mid X = x_i]) = x_i^T \beta$.

So if we can describe $\theta = f_1(\mu(x)), \cdots, f_k(\mu(x)) = f_1(g^{-1}(x^T\beta)), \cdots, f_k(g^{-1}(x^T\beta))$, we can rewrite $f_\theta(z_i) \to f_\beta(y_i \mid x_i, \beta)$. Now we can maximize the likelihood across all observations wrt β to get our estimator for $\hat{\beta}$![2]

The problem, of course, is that this is not likely to have a closed-form solution for β and would likely involved stochastic gradient ascent, iteratively reweighted least squares, Fisher's scoring method, Bayesian methods with approximations of the posterior distribution, etc.

12.1.1 Canonical Exponential Families, Canonical Links, and an example.

Often, we analyze the case where $k = 1$. We can write many such distributions as a *canonical exponential family*, of the form:

$$f_\theta(z) = \exp\left(\frac{z\theta - b(\theta)}{\phi} + c(z, \phi)\right)$$

where ϕ is called the *dispersion parameter* (and as usual, for our purposes, we can substitute $Z = Y \mid X_i$ where X_i affects θ). If ϕ is unknown, we may be dealing with $k = 2$ (e.g. for Normal distribution, $\phi = \sigma^2$, so if σ^2 (and μ) aren't known, $k = 2$); if ϕ is known, you'll have a one-parameter exponential family with θ as the *canonical parameter*.

You may often hear about the *shape* and *scale* of a distribution. These can be thought of as directly related to θ and ϕ respectively: $f_\theta(\frac{x}{\phi})$.

Recall that for the log-likelihood function $l(\theta)$, we have:

[2] Excitement, not factorial.

$$E\left[\frac{\partial l}{\partial \theta}\right] = 0$$

$$E\left[\frac{\partial^2 l}{\partial \theta^2}\right] + E\left[\frac{\partial l}{\partial \theta}\right]^2 = 0 \text{ (Fisher Information equivalence)}$$

We can use this to show that for a canonical exponential family:

$$E[Z] = b'(\theta)$$

$$Var(Z) = \phi \times b''(\theta)$$

For our purposes, we consider $Z = Y \mid X_i$, so $E[Z] = E[Y \mid X] = \mu(X) = b'(\theta)$.

Regarding link functions, given the form we've described at first, it would be nice to have a map g that maps μ to the canonical parameter θ, i.e., $g(\mu) = \theta$. This is called the *canonical link function for the exponential family*. But we have $\mu = b'(\theta)$, so $g(b'(\theta)) = \theta \implies g = b'^{-1}$ – the link function is the functional inverse of $b(\theta)$.

12.2 A GLM workflow.

All of this might feel like gobbledygook. Let's describe the general flow of how you'd connect everything together. We are given n iid samples (X_i, Y_i) (assume we've augmented each X_i by prepending it with a 1, to capture a potential offset):

1. Choose a model $Y \mid X \sim ExpFamilyMember(\pi)$ for model parameter π. We assume π depends on a linear combination of the components of X through a fixed β, i.e. through $X^T\beta$ (so you can imagine write $\pi(X;\beta)$ or π_X everywhere, but the notation is suppressed for convenience). This is why this is a generalized *linear* model.

2. Decide on a link function $g : Dom(\mu(x)) \to \mathbb{R}$ so that $g(\mu(X)) = X^T\beta$ (for the β which you are aiming to estimate).

   If you would like to use the canonical link function $g(\mu) = \theta = H(\pi)$, get the likelihood function in canonical form to determine the form of $H(\pi)$.

3. Determine the relationship $\mu(X) = E[Y \mid X] = f(\pi)$.

   If using the canonical exponential family, it will turn out that $\mu(x) = f(\pi) = b'(\theta) = b'(H(\pi))$.

4. Invert $g(\mu(X)) = X^T\beta = g(f(\pi)) \to \pi = f^{-1}(g^{-1}(X^T\beta)) = h(X^T\beta)$.

   If using the canonical link function, you already fixed that $g(\mu(x)) = H(\pi)$, so $f^{-1}(g^{-1}(\cdot)) = h(\cdot) = H^{-1}(\cdot)$.

5. Replace $\pi \to h(X^T\beta)$ in your likelihood function, and maximize likelihood wrt β to get an estimator $\hat{\beta}(X)$. This likelihood function will likely not be analytically solvable, so you may need to collect a sample and plug in the realized values $X = x$, then use stochastic gradient ascent, iteratively reweighted least squares, etc. to approximate $\hat{\beta}(x)$. (These methods were not discussed in this course.)

6. Bonus: If the model is well-specified (with true model parameter π^*, described with true parameter β^*) and the MLE conditions are met, you have asymptotic Normality: $\sqrt{n}(\hat{\beta} - \beta^*) \xrightarrow[n\to\infty]{(d)} N(0, \mathcal{I}(\beta^*)^{-1})$. That means you can perform hypothesis tests on β^*, π^*, etc. using e.g. Delta Method.

You noticed that many things became "fixed" and directly related to functions in the exponential family's form if we decided to use the canonical link function. This is why many people refer to the canonical link as a "natural" choice that leads to "natural" parameters to study.

12.2.1 Logit link function, and logistic ~~classification~~ regression.

One of people's favorite link functions is the *logit link function*:

$$g(\mu) = \log\left(\frac{\mu(X)}{1-\mu(X)}\right) = X^T\beta$$

This is the canonical link function for a Bernoulli distribution (which is an incredibly important distribution, e.g. can model probabilities of events occurring):

$$\begin{aligned} f_p(y \mid x) &= p^y(1-p)^{1-y} \\ &= \exp\left(\ln\left(p^y(1-p)^{1-y}\right)\right) \\ &= \exp(y\ln(p) + (1-y)\ln(1-p)) \\ &= \exp(y(\ln(p) - ln(1-p)) + \ln(1-p)) \\ &= \exp\left(y\ln\left(\frac{p}{1-p}\right) + \ln(1-p)\right) \end{aligned}$$

The above means that $\theta(p) = \ln\left(\frac{p}{1-p}\right)$, and for Bernoulli distribution, $\mu(x) = E[Y \mid X = x] = p$.

In fact, people love using the (canonical) logit link function for a Bernoulli model so much that if you:

1. invert the function and plug in appropriately:

$$\theta(\mu) = \ln\left(\frac{\mu}{1-\mu}\right) = g(\mu) = X^T\beta \to \mu(\theta) = \exp\left(\frac{1}{1+e^{-\theta}}\right) = \exp\left(\frac{1}{1+e^{-X^T\beta}}\right) = C(X)$$

2. choose an arbitrary threshold t,
3. use C and t as a classifier (classify all X such that $C(X) \geq t$ as, say, $+1$, and the rest as -1)

then you'll have made the classifier known as *logistic regression*.[3]
The Wikipedia article on GLMs has more links and useful information about GLMs.

[3] Apparently they were so enamored by the process that they didn't notice that they end up performing *classification*, not regression.