Fundamentals of Statistics

Manish Agarwal

Oct 2022

These notes are based on course MITx18.6501x 'Fundamentals of Statistics'.

Contents

1 Probability results **2**
1.1 LLT and CLT . . . 2
1.2 Convergence . . . 2
1.3 Slutsky and Continuous Mapping Theorem . . . 3

2 Foundations of Inference **4**
2.1 Bias and Variance of an estimator . . . 4
2.2 Confidence Intervals . . . 5
2.3 Multidimensional Random Variable . . . 7

3 Methods of Estimation **9**
3.1 Maximum Likelihood estimation . . . 9
3.2 The method of Moments . . . 11
3.3 M-estimation . . . 12

4 Hypothesis testing **16**
4.1 Parametric Hypothesis testing . . . 16
4.2 Multiple Hypothesis Testing . . . 19
4.3 Nonparametric Hypothesis testing . . . 19

5 Bayesian Statistics **22**
5.1 Prior and Posterior . . . 22
5.2 Choosing the prior . . . 22
5.3 Inference . . . 23

6 Linear Models **25**
6.1 Linear Regression . . . 25
6.2 Generalized Linear Models . . . 27
6.3 Principle Component Regression . . . 29

1 Probability results

Let $X_1, \ldots, X_n$ be i.i.d. random variables. The fundamental flow of data modelling follows

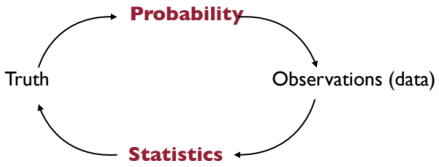

Figure 1: The flow in data science. The true model generates a probability, which are reflected in the observations. Statistics uses this data to infer about the truth that generated it.

1.1 LLT and CLT

Averages of random variables occur naturally in statistics. We make modelling assumptions to apply probability results. The most important results from probability are the following two asymptotic laws.

- **LLN**: Weak and Strong law of Large numbers impose **consistency**

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i \xrightarrow[n\to\infty]{\boldsymbol{P},\ a.s.} \boldsymbol{E}[X] := \mu$$

- **CLT**: Central limit Theorem (rule of thumb $n \geq 30$) imposes **asymptotic distribution**

$$\sqrt{n}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \sigma^2).$$

If n is not large enough to apply CLT we can use

Theorem 1.1. *(**Hoeffding's inequality**) Let n be a finite positive integer and $X_1, \ldots, X_n$ be iid random variables such that $\mu = \boldsymbol{E}[X]$ and $X \in [a, b]$ almost surely, then:*

$$\boldsymbol{P}[|\bar{X}_n - \mu| \geq \epsilon] \leq 2e^{-\frac{2n\epsilon^2}{(b-a)^2}},\ \forall \epsilon > 0$$

Because of CLT, the Gaussian/Normal distribution is ubiquitous in statistics, where the tails decay very fast, almost in finite interval. The pdf is given by $f_{\mu,\sigma^2}(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{1}{2\sigma^2}(x-\mu)^2}$. The standard normal $Z = \frac{X-\mu}{\sigma}$ is the most important building block in statistics. Quantiles show up in confidence intervals and are defined as follows:

Definition 1.1. *(**Quantile**) Let α be in $(0, 1)$. The quantile of order $1-\alpha$ of a random variable X is the number q_α such that $\boldsymbol{P}[X \leq q_\alpha] = 1 - \alpha$.*

1.2 Convergence

Thus if F is the cdf $F(q_\alpha) = 1-\alpha$. With F invertible $q_\alpha = F^{-1}(1-\alpha)$. Also, $\boldsymbol{P}[X > q_\alpha] = \alpha$. If $X = Z \sim \mathcal{N}(0, 1)$ then $\boldsymbol{P}[|X| > q_{\frac{\alpha}{2}}] = \alpha$. The 5% confidence interval corresponding to 1.96σ for two sided distribution.

An important idea to appreciate is the types of convergence of random variable series. If $(T_n)_{n\geq 1}$ is a sequence of random variables then here are the three most important types of convergence

- **Almost surely (a.s.) convergence**:

$$T_n \xrightarrow[n\to\infty]{a.s.} T \iff \boldsymbol{P}\left[\left\{\omega : T_n(\omega) \xrightarrow[n\to\infty]{} T(\omega)\right\}\right] = 1$$

- **Convergence in probability**:

$$T_n \xrightarrow[n\to\infty]{\boldsymbol{P}} T \iff \boldsymbol{P}\left[|T_n - T| \geq \epsilon\right] \xrightarrow[n\to\infty]{} 0, \ \forall \epsilon > 0.$$

- **Convergence in distribution**:

$$T_n \xrightarrow[n\to\infty]{(d)} T \iff \boldsymbol{E}\left[f(T_n)\right] \xrightarrow[n\to\infty]{} \boldsymbol{E}\left[f(T)\right]$$

  for all continuous and bounded function f.

If $(T_n)_{n\geq 1}$ converges a.s. then it also converges in probability, and the two limits are equal a.s. If $(T_n)_{n\geq 1}$ converges in probability, then it also converges in distribution. Convergence in distribution does not imply convergence of probabilities, in general, except when the limit has a density, e.g. Gaussian: $T_n[n \to \infty](d) \implies \boldsymbol{P}[a \leq TN \leq b] \xrightarrow[n\to\infty]{} \boldsymbol{P}[a \leq T \leq b]$.

1.3 Slutsky and Continuous Mapping Theorem

Another important tool in our arsenal is

Theorem 1.2. *(**Slutsky's Theorem**) Let $(X_n), (Y_n)$ be two sequences of r.v., such that $T_n \xrightarrow[n\to\infty]{(d)} T$, and $U_n \xrightarrow[n\to\infty]{\boldsymbol{P}} u$ where T is a r.v. and u is a given real number with deterministic limit $\boldsymbol{P}[U = u] = 1$. Then,*

- $T_n + U_n \xrightarrow[n\to\infty]{(d)} T + u$,
- $T_n U_n \xrightarrow[n\to\infty]{(d)} Tu$,
- *if $u \neq 0$, then $\frac{T_n}{U_n} \xrightarrow[n\to\infty]{(d)} \frac{T}{u}$.*

Finally to work with any general continuous function we make use of

Theorem 1.3. *(**Continuous Mapping Theorem**) If f is a continuous function then*

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

We need f to be continuous around T.

The trinity of classical statistical inference is estimation, confidence intervals, and hypothesis testing. We will tackle them one by one.

2 Foundations of Inference

Definition 2.1. *(**Statistical Model**) Let the observed outcome of a statistical experiment be a sample $X_1, \dots, X_n$ of n iid random variables in some measurable space E, usually $E \subseteq \boldsymbol{R}$, and denote by $\boldsymbol{P}$ their common distribution. A statistical model associated to that statistical experiment is a pair*

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

where

- *E is called the sample space.*
- *$(\boldsymbol{P}_\theta)_{\theta \in \Theta}$ is a family of probability measures on E*
- *Θ is an y set, called the parameter set.*

We will assume that the probability model is **well specified**, i.e. defined such that $\boldsymbol{P} = \boldsymbol{P}_\theta$, for some $\theta \in \Theta$. This particular θ is called the **true parameter**, and is unknown. The aim of the statistical experiment is to **estimate** θ and related relationships. We often assume $\Theta \subseteq \boldsymbol{R}^d$ for some $d \geq 1$: this model is called **parametric**. Sometimes we could have Θ be infinite dimensional in which case the model is called **nonparametric**. If $\Theta = \Theta_1 \times \Theta_2$ where Θ_1 is finite dimensional and Θ_2 is infinite dimensional we call it **semiparametric**. In these models we only care to estimate the finite dimensional parameter and the infinite dimensional one is called nuisance parameter. For example,

- If $X_1, \dots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \boldsymbol{R}$ and $\sigma^2 > 0$ then the **density estimation** statistical model is $(\boldsymbol{R}, (\mathcal{N}(\mu, \sigma^2))_{(\mu,\sigma^2) \in (\boldsymbol{R} \times (0,\infty))})$
- If $(X_1, , Y_1), \dots, (X_n, Y_n) \in \boldsymbol{R}^d \times \boldsymbol{R}$ are iid from the **linear regression model** $Y_i = \beta^T X_i + \epsilon_i$, $\epsilon \overset{iid}{\sim} \mathcal{N}(0, 1)$ for an unknown $\beta \in \boldsymbol{R}^d$ and $X_i \sim \mathcal{N}_d(0, \boldsymbol{I}_d)$ independent of ϵ_i then the statistical model has $E = \boldsymbol{R}^d \times \boldsymbol{R}$ and $\Theta = \boldsymbol{R}^d$.

Definition 2.2. *(**Identifiability**) The parameter θ is called identifiable iff the map $\theta \in \Theta \mapsto \boldsymbol{P}_\theta$ is injective, i.e., $\theta \neq \theta' \implies \boldsymbol{P}_\theta \neq \boldsymbol{P}_{\theta'}$ or equivalently, $\boldsymbol{P} = \boldsymbol{P}_{\theta'} \implies = \theta'$.*

An example of non-identifiable model is $X_i = \mathbf{1}_{Y_i \geq 0}$, called indicator function, for $Y_1, \dots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, for some unknown $\mu \in \boldsymbol{R}$ and $\sigma^2 > 0$, are unobserved. Here μ and σ are not identifiable, but μ/σ is.

Definition 2.3. *(Statistic, estimator, consistency, asymptotic normality)*

- ***Statistic**: Any measurable function of the sample, e.g., $\bar{X}_n, \max_i X_i$, sample variance, etc.*
- ***Estimator** of θ: Any statistic whose expression does not depend on θ.*
- *An estimator $\hat{\theta}_n$ of θ is **weakly consistent** if $\hat{\theta}_n \xrightarrow[n \to \infty]{P} \theta$ w.r.t $\boldsymbol{P}_\theta$, and **strongly consistent** if $\hat{\theta}_n \xrightarrow[n \to \infty]{a.s.} \theta$ w.r.t $\boldsymbol{P}_\theta$.*
- *An estimator $\hat{\theta}$ of θ is **asymptotically normal** if $\sqrt{n}(\hat{\theta}_n - \theta)[n \to \infty](d)\mathcal{N}(0, \sigma^2)$. The quantity σ^2 is then called **asymptotic variance** of $\hat{\theta}_n$.*

2.1 Bias and Variance of an estimator

Definition 2.4. *(**Bias of an estimator**) The bias of an estimator $\hat{\theta}_n$ of θ is given by*

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

*When $bias(\hat{\theta}) = 0$ we say the estimator is **unbiased**.*

Definition 2.5. *(**Variance of an estimator**) An estimator is a random variable so have a variance*

$$V[\hat{\theta}_n] = \boldsymbol{E}[\hat{\theta}_n^2] - (\boldsymbol{E}[\hat{\theta}_n])^2$$

We want estimators to have low bias and low variance at the same time.

Definition 2.6. *(**Quadratic risk**) The risk of an estimator $\hat{\theta}_n \in \boldsymbol{R}$ is*

$$R(\hat{\theta}_n) = \boldsymbol{E}[|\hat{\theta}_n - \theta|^2] = V[\hat{\theta}_n] + bias(\hat{\theta}_n)^2$$

Low quadratic risk means that both bias and variance are small.

2.2 Confidence Intervals

Definition 2.7. *(**Confidence Intervals**: non-asymptotic and asymptotic) Let $(E, (\boldsymbol{P}_\theta)_{\theta\in\Theta})$ be a statistical model based on observations $X_1, \ldots, X_n$ and assume $\Theta \subseteq \boldsymbol{R}$. Let $\alpha \in (0,1)$. The confidence interval (C.I.) of level $1-\alpha$ for θ is any random interval $\mathcal{I}$ depending on $X_1, \ldots, X_n$ whose boundaries do not depend on θ and such that*

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

$\mathcal{I}\ni\theta$ means that $\mathcal{I}$ contains θ. This notation emphasizes the randomness of $\mathcal{I}$. C.I. of asymptotic level $1-\alpha$ for θ is any random interval $\mathcal{I}$ whose boundaries do not depend on θ and such that

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

Example 2.1. *For data $R_1, \ldots, R_n \overset{iid}{\sim} Ber(p)$ for some unknown $p \in (0,1)$ we have the statistical model $\Big(\{0,1\}, (Ber(p))_{p\in(0,1)}\Big)$. We investigate the estimator of p, $\hat{p} = \bar{R}_n = \frac{1}{n}\sum_{i=1}^{n} R_i$. From CLT*

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

This means the cdf $\Phi(x)$ of $\mathcal{N}(0,1)$ is approximately equal to the cdf $\Phi_n(x)$ of $\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{p(1-p)}}$ as n become large. Hence for all $x > 0$, $\boldsymbol{P}[|\bar{R}_n - p| \geq x] \approx 2\left(1 - \Phi\left(\frac{x\sqrt{n}}{\sqrt{p(1-p)}}\right)\right)$. Now for a fixed $\alpha \in (0,1)$, if $q_{\alpha/2}$ is the $(1-\alpha/2)$-quantile of $\mathcal{N}(0,1)$, then with probability $\approx 1-\alpha$, if n is large enough,

$$\bar{R}_n \in \left[p - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, p + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right],$$

yields

$$\lim_{n\to\infty} P\left[\left(\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}\right) \ni p\right] = 1-\alpha$$

*This is **not** a confidence interval because it depends on p.* □

There are three ways to go about it.

- Conservative bound: If we can find a conservative bound, e.g., $p(1-p) \leq \frac{1}{4}$ in this example, we can plug it in and get $\mathcal{I}_{conserv}$. In this example we have

  $$\mathcal{I}_{conserv} = \left[\bar{R}_n - \frac{q_{\alpha/2}}{2\sqrt{n}}, \bar{R}_n + \frac{q_{\alpha/2}}{2\sqrt{n}}\right]$$

  This also is the asymptotic confidence interval since $\lim_{n\to\infty} \boldsymbol{P}[\mathcal{I}_{conserv} \ni p] \geq 1-\alpha$.

- Solving the equation for p: If possible we can also solve the system of two inequalities for the parameter (p in our example)

  $$\bar{R}_n - \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}} \leq p \leq \bar{R}_n + \frac{q_{\alpha/2}\sqrt{p(1-p)}}{\sqrt{n}}$$

  This leads to $(p-_n)^2 \leq \frac{q^2_{\alpha/2}p(1-p)}{n}$ and we need to find the roots $p_1 < p_2$ of the resulting equation leading to the new confidence interval $\mathcal{I}_{solve} = [p_1, p_2]$ such that $\lim_{n\to\infty} \boldsymbol{P}[\mathcal{I}_{solve} \ni p] = 1-\alpha$

- plug-in estimates: Recall by LLN $\hat{p} = \bar{R}_n[n \to \infty]\boldsymbol{P}, a.s.p$, so by Slutsky's theorem we also have $\sqrt{n}\frac{\bar{R}_n - p}{\sqrt{\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$. This leads to

  $$\mathcal{I}_{plug-in} = \left[\bar{R}_n - \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}, \hat{R}_n + \frac{q_{\alpha/2}\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}}\right]$$

  such that $\lim_{n\to\infty} \boldsymbol{P}[\mathcal{I}_{plug-in} \ni p] = 1-\alpha$.

If one uses the exact distribution of $n\bar{R}_n$ one can get finite sample intervals.

When the interest is in some continuous function of the average we use **the delta method**. Let $(Z_n)_{n\geq 1}$ be a sequence of random variable that satisfies $\sqrt{n}(Z_n-\theta) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,\sigma^2)$ for some $\theta \in \boldsymbol{R}$ and $\sigma^2 > 0$. Let $g : \boldsymbol{R} \to \boldsymbol{R}$ be continuously differentiable at that point θ. Then $g((Z_n)_{n\geq 1})$ is also asymptotically normal around $g(\theta)$ and

$$\sqrt{n}\left(g(Z_n)-g(\theta)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,(g'(\theta))^2\sigma^2).$$

The real interpretation of CI is that if we were to repeat the experiment then θ would be in the resulting confidence interval about $(1-\alpha)\%$ of the time.

Example 2.2. *Consider a sample n, iid continuous random variables $X_1,\ldots,X_n$ with density $f(x)=e^{-(x-a)}\mathbf{1}_{x\geq a}$ where $a \in \boldsymbol{R}$ is an unknown parameter. We want to find an asymptotic and non-asymptotic estimator of a, its confidence interval and compare the two.*

***Asymptotic estimator**: We first consider the mean of the data $\frac{1}{n}\sum_{i=1}^{n} X_i$ with the anticipation that it has some information about a. By LLN we have,*

$$\bar{X}_n \xrightarrow[n\to\infty]{\boldsymbol{P}} \boldsymbol{E}[X_1]$$

We can easily see that $\boldsymbol{E}[X_1]=a+1$. Thus, we can thus consider the estimator

$$\hat{a}_1 = \frac{1}{n}\sum_{i=1}^{n} X_i - 1$$

We also can easily calculate $E[X_1^2]=a^2+2(a+1)$ giving $Var[X_1]=1$. We now check the asymptotic normality of this estimator. By CLT

$$\sqrt{n}\left(\bar{X}_n-\boldsymbol{E}[X_1]\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0,Var[X_1]\right) \implies \sqrt{n}\left(\hat{a}_1-a\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0,1\right)$$

To find the confidence interval $\mathcal{I}_1=\hat{a}_1+[-s,s]$ with $s>0$ with confidence level $(1-\alpha)$ we look at the coverage condition. We want $\boldsymbol{P}[\mathcal{I}_1 \ni a] \xrightarrow[n\to\infty]{(d)} \boldsymbol{P}[Z\in[-q,q]]=1-\alpha=2\Phi(q)-1$, where Z is a standard normal. Now,

$$\mathcal{I}_1\ni a \iff a\in[\hat{a}_1-s,\hat{a}_1+2] \iff \hat{a}_1-s\leq a\leq \hat{a}_1+s \iff -s\leq \hat{a}_1-a\leq s \iff -\sqrt{n}s\leq\sqrt{n}(\hat{a}_1-a)\leq\sqrt{n}s$$

We define $q=\sqrt{n}s$. We note that $q=q_{\alpha/2}$, i.e., the $1-\frac{\alpha}{2}$ quantile for a two tailed confidence interval. Thus, $\mathcal{I}_1=\hat{a}_1+\left[-\frac{q_{\alpha/2}}{\sqrt{n}},\frac{q_{\alpha/2}}{\sqrt{n}}\right]$.

***Non-asymptotic estimator**: We notice that the parameter is the boundary of the exponential and hence a plausible estimator is*

$$\hat{a}_2 = \min_{1\leq i\leq n} X_i$$

This will be doable since a CDF of a continuous variable has enough information about everything we need here. For $t<a$ we have $\boldsymbol{P}[\hat{a}_2>t]=1$. For $t>a$, we have $P[\hat{a}_2>t]=\boldsymbol{P}[X_i>t,\forall i]=\boldsymbol{P}[\bigcap_{i=1}^{n}\{X_i>t\}]=\prod_{i=1}^{n}\boldsymbol{P}[X_i>t]=e^{-n(t-a)}$. With $t=\frac{1}{n}\tilde{t}+a$ we have $[\hat{a}_2>\frac{1}{n}\tilde{t}+n]=e^{-\tilde{t}}$. Thus the non-asymptotic distribution is,

$$n(\hat{a}_2-a)\sim Exp(1)$$

We intend to estimate $\mathcal{I}_2=[\hat{a}_2-s,\hat{a}_2]$, this one sided interval is justified as the boundary can only be to the left of hata$_2$. We want $\boldsymbol{P}[\mathcal{I}_2\ni a]=1-\alpha$. Now,

$$\mathcal{I}_2\ni a \iff a\in[\hat{a}_2-s,\hat{a}_2] \iff \hat{a}_2-s\leq a\leq\hat{a}_2 \iff \hat{a}_2-a\leq s \iff n(\hat{a}_2-a)\leq ns$$

we consider $q=ns$. Further, $\boldsymbol{P}[\mathcal{I}_2\ni a]=\boldsymbol{P}[Y\leq q]$ where $Y\sim Exp(1)$. Thus, $\boldsymbol{P}[\mathcal{I}_2\ni a]=1-e^{-q}=1-\alpha \implies q=\log\frac{1}{\alpha}$, giving $\mathcal{I}_2=\left[\hat{a}_2=\frac{1}{n}\log\frac{1}{\alpha},\hat{a}_2\right]$.

We note that the non-asymptotic confidence interval is much tighter than the asymptotic confidence interval. □

2.3 Multidimensional Random Variable

For the multivariate case, let $f : \boldsymbol{R}^d \to \boldsymbol{R}$ be the function of interest with gradient $\nabla f : \boldsymbol{R}^d \to \boldsymbol{R}^d$ and the Hessian Matrix $\nabla^2 f := \boldsymbol{H} f : \boldsymbol{R}^d \to \boldsymbol{R}^{d\times d}$, with $(\boldsymbol{H}f)_{ij} := \frac{\partial^2}{\partial\theta_i\partial\theta_j} f, 1 \le i,j \le d$. If the Hessian Matrix is negative semi-definite the function f is concave and has a unique maxima. A multivariate random variable, is a vector-valued function whose components are random variables on the same underlying probability space with $X :\to \boldsymbol{R}^d$, where each $X^{(k)}$ is a random variable on Ω. The probability distribution of a random vector X is the joint distribution of its components $X^{(1)}, \dots, X^{(d)}$. The cumulative distribution function of a random vector X is defined as $F : \boldsymbol{R}^d \to [0,1]$ as $x \mapsto \boldsymbol{P}[X^{(1)} \le x^{(1)}, \dots, X^{(d)} \le x^{(d)}]$.

Convergence in Probability in higher dimension: To make sense of the consistency statement $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\boldsymbol{P}} \theta^*$ where the MLE $\hat{\theta}_n^{MLE}$ is a random vector, we need to know what convergence in probability means in higher dimensions. But this is no more than the convergence in probability in each component. Let $X_1, X_n, \dots$ be a sequence of random vectors of size $d \times 1$, i.e., $X_i = \begin{bmatrix} X_i^{(1)} \\ \vdots \\ X_i^{(d)} \end{bmatrix}$. Let $X = \begin{bmatrix} X^{(1)} \\ \vdots \\ X^{(d)} \end{bmatrix}$ be another vector of size $d \times 1$. Then $X_n \xrightarrow[n\to\infty]{\boldsymbol{P}} X \iff X_n^{(k)} \xrightarrow[n\to\infty]{\boldsymbol{P}} X^{(k)}, \ \forall 1 \le k \le d$.

Covariance matrix In higher dimensions one examines the correlation between different components beyond the variance of individual components $Cov[X,Y] = \boldsymbol{E}[XY] - \boldsymbol{E}[X]\boldsymbol{E}[Y]$. If X and Y are independent, then $Cov[X,Y] = 0$. In general, the converse is not true expect if $(X,Y)^T$ is a Gaussian vector, i.e. $\alpha X + \beta Y$ is Gaussian for all $(\alpha,\beta) \in \boldsymbol{R}^2/\{0,0\}$. As an example, take $X \sim \mathcal{N}(0,1)$, $B \sim Ber(\frac{1}{2})$, then we have the Rademacher variable $R = 2B - 1 \sim Rad(\frac{1}{2}$. Then $Y = RX \sim \mathcal{N}(0,1)$. However taking $\alpha = \beta = 1$ we get $X + Y = \begin{cases} 2X & prob\ \frac{1}{2} \\ 0 & prob\ \frac{1}{2} \end{cases}$, conditionally on X. Actually, $Cov[X,Y] = 0$ but they are not independent $|X| = |Y|$. Also, $Cov[X]$ is a $d \times d$ matrix, with $Cov[AX + B] = ACov[X]A^T$. A multivariate Gaussian pdf is given by

$$f_X(x) = \frac{1}{\sqrt{(2\pi)^d \boldsymbol{det}(\Sigma)}} e^{-\frac{1}{2}(x-\mu)^T\Sigma^{-1}(x-\mu)}, x \in \boldsymbol{R}^d.$$

Multivariate CLT: Let X be a random variable of dimension d and let μ and Σ be its mean and covariance. Let $X_1, \dots, X_n$ be iid copies of X. Let $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. Then

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

or equivalently

$$\sqrt{n}\Sigma^{-1/2}(\bar{X}_n - \mu) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, I_d)$$

Multivariate Delta Method: The multivariate delta method states that given a sequence of random vectors $(T_n)_{n\ge1}$ satisfying $\sqrt{n}(T_n - \theta) \xrightarrow[n\to\infty]{(d)} T$, a function $g : \boldsymbol{R}^d \to \boldsymbol{R}^k$ that is continuously differentiable at θ, then

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \nabla g(\theta)^T T$$

For $Tn = \bar{X}_n$ and $\theta = \boldsymbol{E}[X]$, CLT gives $T \sim \mathcal{N}(0,\Sigma_X)$. For this case we have

$$\sqrt{n}\,(g(T_n) - g(\theta)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, \nabla g(\theta)^T \Sigma_X \nabla g(\theta))$$

Example 2.3. ***Asymptotic distribution of sample median***: *Now we derive the asymptotic variance of the sample median. Let m_n be the sample median and suppose that we observe data $X_1, \dots, X_n \sim F$, where F is a continuous distribution, and let μ bet he population median $F(\mu) = \frac{1}{2}$. For the following assume that n is odd, so that the sample median is unique point $m_n = X_{\frac{n+1}{2}}$. Then we will show that $\sqrt{n}(m_n - \mu) \le a \xrightarrow[n\to\infty]{P} \mathcal{N}(0,\sigma^2)$, for some asymptotic variance σ^2. To this end, notice that*

$$\left\{ m_n \le \mu + \frac{a}{\sqrt{n}} \right\} = \left\{ \#(X_i \le \mu + \frac{a}{\sqrt{n}}) \ge \frac{n+1}{2} \right\} = \left\{ \bar{Y}_n \ge \frac{n+1}{2} \right\},$$

where the random variables Y_i are defined by $Y_i = \mathbf{1}_{X_i \le \mu + \frac{a}{\sqrt{n}}}$. Then we have

$$\begin{aligned} P(\sqrt{n}(m_n - \mu) \le a) &= P\left(\bar{Y}_n \ge \frac{n+1}{2}\right) \\ &= P\left(\bar{Y}_n - np_n \ge \frac{n+1}{2} - np_n\right) \\ &= P\left(\frac{\bar{Y}_n - np_n}{\sqrt{np_n(1-p_n)}} \ge \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right) \end{aligned}$$

where $p_n = F(\mu + \frac{a}{\sqrt{n}})$ is the probability that $Y_i = 1$, notice that Y_i is nothing more than a Bernoulli(P_n) random variable. Let $p = F(\mu) = \frac{1}{2}$. Notice that

$$\frac{\bar{Y}_n - np_n}{\sqrt{np_n(1-p_n)}} - \frac{\bar{Y}_n - np}{\sqrt{np(1-p)}} \xrightarrow[n\to\infty]{P} 0,$$

which then yields

$$\frac{\bar{Y}_n - np_n}{\sqrt{np_n(1-p_n)}} \xrightarrow[n\to\infty]{d} \mathcal{N}(0,1)$$

After some manipulation we get

$$P(m_n \le \mu + \frac{a}{\sqrt{n}}) = P\left(Z \ge \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}}\right).$$

We finally notice that

$$\begin{aligned} \frac{\frac{n+1}{2} - np_n}{\sqrt{np_n(1-p_n)}} &= \frac{\frac{n}{2} + \frac{1}{2} - nF(\mu + \frac{a}{\sqrt{n}})}{\sqrt{np_n(1-p_n)}} = \frac{\frac{n}{2} - nF(\mu + \frac{a}{\sqrt{n}})}{\sqrt{np_n(1-p_n)}} + \frac{\frac{1}{2}}{\sqrt{np_n(1-p_n)}} \\ &= \frac{F(\mu) - F(\mu + \frac{a}{\sqrt{n}})}{\frac{a}{\sqrt{n}}} \frac{a}{\sqrt{np_n(1-p_n)}} + \frac{\frac{1}{2}}{\sqrt{np_n(1-p_n)}} \\ &\xrightarrow[n\to\infty]{d} -2aF'(\mu) = -2af(\mu) \end{aligned}$$

Therefore

$$\sqrt{n}(m_n - \mu) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \frac{1}{4f(\mu)^2}).$$

For a normal distribution the asymptotic variance of sample mean is σ^2, while for sample median is $\frac{\pi}{2}\sigma^2$ □

To see this more generally, suppose $X_1, \ldots, X_n$ are iid continuous r.v. from distribution with cdf F_X. Let $Y_n(x) = \frac{1}{n}\sum_{i=1}^{n} \mathbf{1}_{X_i \le x} = \frac{1}{n}\sum_{i=1}^{n} Z_i$, where $Z_i(x)$ is the indicator function. Then, Z_i has expectation $\mu(z) = F_X(z)$ and variance $\sigma^2(x) = F_X(x)(1 - F_X(x))$, and by CLT

$$\sqrt{n}(Y_n(x) - F_X(x)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, F_X(x)(1 - F_X(x))).$$

Now for a transformation though a function $g(t) = F_X^{-1}(t)$ defined for $0 < t < 1$ we have $g'(t) = \frac{1}{f_X(F_X^{-1}(t))}$. Thus, using delta method

$$\sqrt{n}\left(F_X^{-1}(Y_n(x)) - F_X^{-1}(F_X(x)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \frac{F_X(x)(1-F_X(x))}{(f_X(F_X^{-1}(F_X(x))))^2}\right)$$

Now, we recognize $q = F(x)$ as the q-th sample quantile giving

$$\sqrt{n}\left(F_X^{-1}(Y_n(x)) - x\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \frac{p(1-p)}{(f_X(x))^2}\right)$$

Now $F_X^{-1}(Y_n(x))$ is a rv that lies between the $p$the and $(p-1)$th sample quantile, that can be written using via order statistic notation as $X_{[np]}$. In fact, $|X_{[np]} - F_X^{-1}(Y_n(x))| \xrightarrow{a.s.} 0$. Thus,

$$\sqrt{n}\left(X_{[np]} - x\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(0, \frac{p(1-p)}{(f_X(x))^2}\right).$$

3 Methods of Estimation

There are certain principled ways of finding a general estimator in an experiment - maximum likelihood, method of moments, and M-estimators. All methods yield to asymptotic normality under regularity conditions. We will look at these three methods and then contrast them at the end.

3.1 Maximum Likelihood estimation

Let $(E, (\boldsymbol{P}_\theta)_{\in\Theta})$ be a statistical model associated with a sample of iid r.v. $X_1, \ldots, X_n$. Assume that there exists $\theta^* \in \Theta$ such that $X_1 \sim \boldsymbol{P}_{\theta^*}$: θ^* is the true parameter. Our goal then is to find an estimator $\hat{\theta}(X_1, \ldots, X_n)$ such that $\boldsymbol{P}_{\hat{\theta}}$ is close to $\boldsymbol{P}_{\theta^*}$ for the true parameter θ^*. The first attempt to quantify the distance between the two distributions is *total variation distance* between $\boldsymbol{P}_\theta$ and $\boldsymbol{P}_{\theta'}$ defined by $TV(\boldsymbol{P}_\theta, \boldsymbol{P}_{\theta'}) = \max_{A\subset E} |\boldsymbol{P}_\theta(A) - \boldsymbol{P}_{\theta'}|$ This reduces to $\frac{1}{2}\sum_{x\in E} |p_\theta(x) - p_{\theta'}(x)|$ for discrete and $\frac{1}{2}\int |f_\theta(x) - f_{\theta'}(x)|dx$ for continuous distributions. They follow the properties of symmetry, positivity, definiteness, and triangle inequality and hence is a measure of distance. However it is unclear how to build this function empirically $\theta \mapsto \widehat{TV}(\boldsymbol{P}_\theta, \boldsymbol{P}_{\theta'})$ to do the estimation. **KUllback-Leibler (KL) divergence** (or relative entropy) between two probabilities provides a good solution. The KL divergence between two probability measures $\boldsymbol{P}_\theta$ and $\boldsymbol{P}_{\theta'}$ is defined by

$$KL(\boldsymbol{P}_\theta, \boldsymbol{P}_{\theta'}) = \begin{cases} \sum\limits_{x\in E} p_\theta \log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right) & discrete \\ \int\limits_E f_\theta log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right) dx & continuous \end{cases}$$

KL divergence, however is not symmetric and does not follow triangle's inequality in general, thus is not a distance measure, but rater a divergence. However, asymmetry in KL divergence makes it possible to estimate it! The unique minimizer of the function $\theta \mapsto KL(\boldsymbol{P}_{\theta^*}, \boldsymbol{P}_\theta)$ is θ^* as we will see here. We can write

$$KL(\boldsymbol{P}_{\theta^*}, P_\theta) = \boldsymbol{E}_{\theta^*}\left[log\left(\frac{p_{\theta^*}(X)}{p_\theta(X)}\right)\right] = \boldsymbol{E}_{\theta^*}[log\ p_{\theta^*}(X)] - \boldsymbol{E}_{\theta^*}[log\ p_\theta(X)]$$

So to minimize the function $\theta \mapsto KL(\boldsymbol{P}_{\theta^*}, \boldsymbol{P}_\theta)$ with respect to θ we notice the first term is independent of it, and hence, equivalently, we need to maximize $\boldsymbol{E}_{\theta^*}[\log\ p_\theta(X)]$. Moreover, $\boldsymbol{E}_{\theta^*}[h(X)]$ can by estimated using LLN by $\frac{1}{n}\sum\limits_{i=1}^n h(X_i)$. Thus

$$\min_{\theta\in\Theta} \widehat{KL}(\boldsymbol{P}_{\theta^*}, \boldsymbol{P}_\theta) \iff \min_{\theta\in\Theta} -\frac{1}{n}\sum_{i=1}^n \log\ p_\theta(X_i) \iff \max_{\theta\in\Theta} \prod_{i=1}^n p_\theta(X_i)$$

This is the **maximum likelihood principle**.

Definition 3.1. *(**Likelihood**) Let $(E, (\boldsymbol{P}_\theta)_{\theta\in\Theta})$ be a statistical model associated with a sample of iid r.v. $X_1, \ldots, X_n$. If E is discrete with common pdf P_θ then the likelihood of the model is the map $\mathcal{L}_n$ defined as $\mathcal{L}_n : E^n \times \Theta \to \boldsymbol{R} : (x_1, \ldots, x_n, \theta) \mapsto \prod\limits_{i=1}^n P_\theta[X_i = x_i]$. If E is continuous with common density of $\boldsymbol{P}_\theta$ being f_θ then $\mathcal{L}_n : E^n \times \Theta \to \boldsymbol{R} : (x_1, \ldots, x_n, \theta) \mapsto \prod\limits_{i=1}^n f_\theta(x_i)$.*

Definition 3.2. *(**Maximum likelihood estimator**) Let $X_1, \ldots, X_n$ be an iid sample associated with a statistical model $(E, (\boldsymbol{P}_\theta)_{\theta\in\Theta})$ and let $\mathcal{L}$ be the corresponding likelihood. Then, the maximum likelihood estimator or θ is defined as*

$$\hat{\theta}_n^{MLE} = argmax_{\theta\in\Theta}\mathcal{L}(X_1, \ldots, X_n, \theta),$$

provided it exists. In practice, we use the maximum log-likelihood estimator using the monotonicity property of the logarithm function.

$$\hat{\theta}_n^{MLE} = argmax_{\theta\in\Theta} \log\ \mathcal{L}(X_1, \ldots, X_n, \theta).$$

A function twice differentiable $h : \Theta \subset \boldsymbol{R} \to \boldsymbol{R}$ is said to be **concave** if its second derivative satisfies $h''(\theta) \leq 0,\ \forall\theta \in \Theta$. Is is strictly concave if $h''(\theta) < 0$. Moreover, h is said to be convex if $-h$ is concave. More generally, a multivariate function $h : \Theta \subset \boldsymbol{R}^d \to \boldsymbol{R}, d \geq 2$, is concave if its Hessian is Negative semidefinite, i.e. $x^T \boldsymbol{H} h(\theta) x \leq 0,\ \forall x \in \boldsymbol{R}^d,\ \theta \in \Theta$. The maxima of a concave functions are unique solution to $h'(\theta^*) = 0$,

and in the multivariate case $\nabla h(\theta) = 0 \in \boldsymbol{R}^d$. Convex optimization is the formal theory to find the solutions numerically, when closed form solutions are not available. Log likelihood functions are concave.

Under mild regularity conditions log-likelihood estimator is **consistent**, we have

$$\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\boldsymbol{P}} \theta^*$$

This is because for all $\theta \in \Theta$ $\frac{1}{n}\log \mathcal{L}(X_1, \ldots, X_n, \theta) \xrightarrow[n\to\infty]{\boldsymbol{P}} c - KL(\boldsymbol{P}_{\theta^*}, \boldsymbol{P}_\theta)$. Moreover, the minimizer of the right-hand side is θ^* if the parameter is identifiable. Technical conditions allow to transfer this convergence to the minimizer.

For **asymptotic normality** we also need to address the case when $\theta \subset \boldsymbol{R}^d, d \geq 2$ where its coordinates are not necessarily independent. Multivariate CLT and delta methods can be used here.

Definition 3.3. *(**Fisher Information**) Define the log-likelihood for one observation as $\ell(\theta) = \log\ \mathcal{L}_1(X, \theta)$, $\theta \in \Theta \subset \boldsymbol{R}^d$. Assume that ℓ is almost surely twice differentiable. Under some regularity conditions, the Fisher information of the statistical model is defined as*

$$I(\theta) = \boldsymbol{E}[\nabla\ell(\theta)\nabla\ell(\theta)^T] - [\nabla\ell(\theta)]\boldsymbol{E}[\nabla\ell(\theta)]^T = -\boldsymbol{E}[\boldsymbol{H}\ell(\theta)]$$

If $\Theta \subset \boldsymbol{R}$, we get $I(\theta) = Var[\ell'(\theta)] = -\boldsymbol{E}[\ell''(\theta)]$.

Another useful form for Fisher information is

$$I(\theta) = \int_{-\infty}^{\infty} \frac{\left(\frac{\partial f_\theta(x)}{\partial\theta}\right)^2}{f_\theta(x)} dx$$

The Fisher information captures the negative of the expected curvature of $\ell(\theta)$. High information means low variance.

Theorem 3.1. *(**Asymptotic consistency and normality of MLE**) Let $\theta^* \in \Theta$ be the true parameter. Assume*

- *The parameter is identifiable.*
- *For all $\theta \in \Theta$ the support of $\boldsymbol{P}_\theta$ does not depend on θ.*
- *θ^* is not on the boundary of Θ.*
- *$I(\theta)$ is invertible in a neighbourhood of θ^*.*
- *A few more technical conditions.*

Then, $\hat{\theta}_n^{MLE}$ satisfies:

- ***Consistency**: $\hat{\theta}_n^{MLE} \xrightarrow[n\to\infty]{\boldsymbol{P}} \theta^*$, wrt $\boldsymbol{P}_{\theta^*}$.*
- ***Asymptotic normality**: $\sqrt{n}(\hat{\theta}_n^{MLE} - \theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}_d(0, I(\theta^*)^{-1})$ wrt $\boldsymbol{P}_{\theta^*}$.*

To see the proof of this convergence, notice that via CLT

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_i'(\theta) - \boldsymbol{E}[\ell'(\theta)]\right] \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, Var[\ell'(\theta)])$$

Now, $Var[\ell'(\theta)] = I(\theta)$, $l_i(\theta) = \log f_\theta(X_i)$, $\boldsymbol{E}[\ell'(\theta^*)] = 0$, $\frac{1}{n}\sum_{i=1}^{n}\ell'(\hat{\theta}) = 0$ (because $\hat{\theta}$ maximizes $\frac{1}{n}\sum_{i=1}^{n}\ell_i(\theta)$). Thus, we can write

$$\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_i'(\hat{\theta}) - \ell_i'(\theta^*)\right] \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta))$$

We know that $\hat{\theta}$ converges to θ^* by Consistency. Thus we can do the following Taylor expansion $\ell_i'(\hat{\theta}) - \ell_i'(\theta^*) \approx (\hat{\theta} - \theta^*)\ell_i''(\theta^*)$. Thus, as $n \to \infty$ we have

$$(\hat{\theta} - \theta^*)\sqrt{n}\left[\frac{1}{n}\sum_{i=1}^{n}\ell_i''(\theta^*)\right] \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta^*))$$

By LLN we have $\frac{1}{n}\sum_{i=1}^{n}\ell_i''(\theta^*) \xrightarrow[n\to\infty]{\boldsymbol{P}} \boldsymbol{E}[\ell_i''(\theta^*)] = -I(\theta^*)$. Thus, $(\hat{\theta}-\theta^*)\sqrt{n}I(\theta^*) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, I(\theta^*))$ using Slutsky's theorem. Thus we have

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

In the case we are not able to solve the maximization problem associated with log-likelihood in closed form, we use iterative numerical algorithm. One of them is **Expectation Maximization**, or the EM algorithm. These are particularly important for mixture models. This does not have any guarantee to converge, but usually work in practice. We have input data $X_1, \ldots, X_n$. We initialize the parameters to some value, and then do the two steps until convergence:

- E-step: Compute the weights $w_i = E[Z_i|X_i]]$ for each observation, depending on the previous steps parameter values.
- M-step: Update the values of parameters, using the maximum log-likelihood solutions, assuming weights from the previous step.

Example 3.1. *Given n iid samples $X_1, \ldots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ with $\mu \in \boldsymbol{R}$ and $\sigma^2 > 0$, we want to estimate $\mu - \sigma$.*

We use the usual MLE estimates $\hat{\mu} = \bar{X}_n$ and $\widehat{\sigma^2} = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. We then construct the function $g(x, y) = x - \sqrt{y}$ and note that $g(\hat{\mu}, \widehat{\sigma^2}) \xrightarrow[n\to\infty]{\boldsymbol{P}} \mu - \sigma$. The MLE has asymptotic normality

$$\sqrt{n}\left(\begin{bmatrix}\hat{\mu}\\ \widehat{\sigma^2}\end{bmatrix} - \begin{bmatrix}\mu\\ \sigma^2\end{bmatrix}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, I(\mu, \sigma^2)^{-1}\right)$$

where $I(\mu, \sigma^2) = -\boldsymbol{E}[\ell''(\mu, \sigma^2)] = \begin{bmatrix}\frac{1}{\sigma^2} & 0\\ 0 & \frac{1}{2\sigma^4}\end{bmatrix}$. Now, $\nabla g(x, y) = \begin{bmatrix}1\\ -\frac{1}{2\sqrt{y}}\end{bmatrix}$. Thus, using delta method we get

$$\sqrt{n}\left(\hat{\mu} - \sqrt{\widehat{\sigma^2}} - (\mu - \sigma)\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, \nabla g^T I^{-1} \nabla g) = \mathcal{N}(0, \frac{3}{2}\sigma^2).$$

□

3.2 The method of Moments

Let $X_1, \ldots, X_n$ be an iid sample associated with a statistical model $(E, (\boldsymbol{P}_\theta)_{\theta\in\Theta})$. Assume that $E \subseteq \boldsymbol{R}$ and $\Theta \subseteq \boldsymbol{R}^d$, for some $d \geq 1$. We define **population moments** as $m_k(\theta) = \boldsymbol{E}_\theta[X^k]$. For many distributions all the moments of X are contained in a single function called **Moment Generating function** (MGF) given by $M_X(t) = [e^{tX}], t \in \boldsymbol{R}$. Given this function one can utilize

$$\boldsymbol{E}[X^k] = M_X^{(k)}(t)\Big|_{t=0}$$

For example the MFG of standard Gaussian is $M_Z(t) = e^{\frac{t^2}{2}}$. Next we define **Sample Moments**, where the kth sample moment is $\hat{m}_k = \frac{1}{n}\sum_{i=1}^{n}X_i^k$, and by LLN we have

$$\hat{m}_k \xrightarrow[n\to\infty]{\boldsymbol{P}} m_k$$

The methods of moment estimator $\hat{\theta}_n \in \boldsymbol{R}^d$ satisfies

$$m_k(\hat{\theta_n}) = \hat{m}_k, \ \ for\ k = 1, \ldots, d$$

Thus for $M : \Theta \to \boldsymbol{R}^d, \theta \mapsto M(\theta) = (m_1(\theta), \ldots, m_d(\theta))$. With M one-to-one we have $\theta = M^{-1}(m_1(\theta), \ldots, m_d(\theta))$. Thus we define the Moments estimator of θ as $\hat{\theta}_n^{MM} = M^{-1}(\hat{m}_1, \ldots, \hat{m}_d)$, provided it exists.

To look at the statistical properties we let $\hat{M} = (\hat{m}_1, \ldots, \hat{m}_d)$ and let $\Sigma(\theta) = Cov_\theta(X_1, X_1^2, \ldots, X_1^d)$ be the covariance matrix of the random vector, which we assume to exist. We also assume M^{-1} is continuously differentiable at $M(\theta)$. This can be extended to more general functions of moments. Let $g_1, \ldots, g_d : E \to \boldsymbol{R}$ be given functions and we define $m_k(\theta) = \boldsymbol{E}_\theta[g_k(X)]$ for all $k = 1, \ldots, d$. Let $\Sigma(\theta) = Cov_\theta(g_1(X_1), \ldots, g_d(X_1))$ be the covariance matrix of the random vector which we assume to exist. We also assume M to be one to one and M^{-1} is continuously differentiable at $M(\theta)$. Applying the CLT and Delta method yields:

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

w.r.t. $\boldsymbol{P}_\theta$ where $\Gamma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta}(M(\theta))\right]^T$ and $\Sigma(\theta) = \left[\frac{\partial M^{-1}}{\partial \theta}(M(\theta))\right]$ Comparing the quadratic risks, MLE is more accurate. MLE still gives good results if model is misspecified. However, sometimes the MLE is intractable but MM is easier due to polynomial equations.

3.3 M-estimation

In M-estimation we try to estimate a parameter by minimizing some, intuitively sound, loss function. We find a function $\rho : E \times \mathcal{M} \to \boldsymbol{R}$ where $\mathcal{M}$ is the set of all possible values for the unknown μ^*, such that $\mathcal{Q}(\mu) := \boldsymbol{E}[\rho(X_t, \mu)]$ achieves its minimum at $\mu = \mu^*$. In univariate case with data X_i the process takes the form

$$\hat{\mu} = argmin_{b \in \boldsymbol{R}} \sum_{i=1}^{n} \rho(x_i - m)$$

for some choice of function ρ. $\rho(x) = x^2$ (same as likelihood for normal distribution) yields sample mean, $\rho(x) = |x|$ (same as likelihood for Laplace distribution) yields sample median, and $\rho(x) = \begin{cases} |x-m|, & |x-m| \geq \delta \\ \frac{(x-m)^2}{2\delta} + \frac{\delta}{2}, & |x-m| < \delta \end{cases}$, for some parameter δ, this is called Huber loss function. For $\alpha \in (0,1)$ fixed if we take $\phi(x, \mu) = C_\alpha(x - \mu)$, for all $x \in \boldsymbol{R}, \mu \in \boldsymbol{R}$, then μ^* is the α-quantile of $\boldsymbol{P}$. The **check function** $C_\alpha(x) = \begin{cases} -(1-\alpha)x & if\ x < 0 \\ \alpha x & if x \geq 0 \end{cases}$ gives us the quantile as the minimizer. MLE is a special case of M-estimator for an appropriate ρ function.

Asymptotic Normality of M-estimator: We define the following quantities:

$$J(\mu) := -\frac{\partial^2 Q(\mu)}{\partial \mu \partial \mu^T} = -\boldsymbol{E}\left[\frac{\partial^2 \rho}{\partial \mu \partial \mu^T}(X_1, \mu)\right]$$

the equality being true under some regularity conditions. Also,

$$K(\mu) := Cov\left[\frac{\partial \rho}{\partial \mu}(X_1, \mu)\right].$$

In the log likelihood case, with $\mu = \theta$ we have $J(\theta) = K(\theta) = I(\theta)$.

Thus, for $\mu^* \in \mathcal{M}$ be the true parameter and we assume

- μ^* is the only minimizer of the function $\mathcal{Q}$
- $J(\mu)$ is invertible for all $\mu \in \mathcal{M}$
- A few more technical conditions

then $\hat{\mu}_n$ satisfies:

$$\hat{\mu}_n \xrightarrow[n\to\infty]{\boldsymbol{P}} \mu^*$$

$$\sqrt{n}(\hat{\mu}_n - \mu^*) \xrightarrow[n\to\infty]{(d)} (0, J(\mu)^{-1} K(\mu^*) J(\mu)^{-1}).$$

MSE bias variance tradeoff The MSE of an estimator is given by

$$mse(\hat{\theta}) = \mathbf{E}(\hat{\theta} - \theta)^2$$

where θ is the true parameter. The following decomposition

$$mse(\hat{\theta}) = var(\hat{\theta}) + (\mathbf{E}\hat{\theta} - \theta)^2$$

is called the bias variance decomposition. For two unbiased estimators $\hat{\theta}_1$ and $\hat{\theta}_2$ the **relative efficiency** of $\hat{\theta}_1$ versus $\hat{\theta}_2$ is given by $\text{eff}(\hat{\theta}_1, \hat{\theta}_2) = \frac{var(\hat{\theta}_1)}{\hat{\theta}_2}$.

Example 3.2. *If $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, then we know $\bar{X}_n$ is an unbiased estimator of μ. We have $mse(\bar{X}_n) = var(\bar{X}_n) = \frac{\sigma^2}{n}$. Now consider the estimator $a\bar{X}_n$ for $a \in (0, \infty)$. Notice this estimator is biased if $a \neq 1$, i.e. $\mathbf{E}(x\bar{X}_n) = a\mu$. We have $mse(a\bar{X}_n) = var(a\bar{X}_n) + bias(a\bar{X}_n)^2$. This is minimized when $a = \frac{\mu^2}{\mu^2 + \frac{1}{n}\sigma^2}$, which is not equal to 1. This gives $mse(a\bar{X}_n) = \frac{\sigma^2}{n}\frac{\mu^2}{\mu^2 + \frac{1}{n}\sigma^2} < \frac{\sigma^2}{n}$. Thus, the estimator is biased, but can reduce the variance!*

Repeating the same process to estimate the variance of the normal with fixed mean 0, we see that the estimator $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2$ has $mse(\hat{\sigma}^2) = \frac{2\sigma^4}{n}$. While if we take the shrinkage estimator $a\hat{\sigma}^2$ and minimize it at $a = \frac{n}{2+n}$ we get MSE of $\frac{2\sigma^4}{2+n}$, which again reduces the variance. □

M-estimation is fairly famous in Machine Learning and is utilizes for robust statistics. Some important distributions other than Normal distribution are:

- **Laplace distribution**: For the Laplace distribution, the likelihood of $X_1, \ldots, X_n$ is $\mathcal{L}(X_1, \ldots, X_n; \mu) = \prod_{i=1}^{n} \frac{1}{2} e^{-|x_i - \mu|}$. The log-likelihood is $\ell(X_1, \ldots, X_n; \mu) = -n\log(2) - \sum_{i=1}^{n} |x_i - \mu|$. Thus, $\hat{\mu}_{MLE} = \min_\mu \sum_{i=1}^{n} |x_i - \mu|$. This is an M-estimator that exactly corresponds to the sample median. Now, the asymptotic variance of the sample mean $avar(\bar{X}_n) = Var(X) = 2$ which we can calculate as follows (taking $\boldsymbol{E}X = 0$):

$$\begin{aligned}
Var(X) = \boldsymbol{E}X^2 &= \int_{-\infty}^{\infty} \frac{x^2}{2} e^{-|x|} dx \\
&= 2\int_0^{\infty} \frac{x^2}{2} e^{-|x|} dx \\
&= 2\left(-\frac{x^2}{2} e^{-x} \Big|_0^{\infty} - \int_0^{\infty} x e^{-x} dx \right) \\
&= 2\left(-\int_0^{\infty} x e^{-x} dx \right) \\
&= 2\left(x e^{-x} \Big|_0^{\infty} + \int_0^{\infty} e^{-x} dx \right) \\
&= 2\left(-e^{-x} \Big|_0^{\infty} \right) = 2
\end{aligned}$$

  In the case of Laplace(μ, 1) distribution, $f(\mu) = \frac{1}{2}$. Thus, the asymptotic variance of the median is $avar(m_n) = 1$, which is smaller than the asymptotic variance of the sample mean!

- **Cauchy distribution**: For $X \sim Cauchy(\mu, 1)$ we have

$$f(x) = \frac{1}{\pi(1 + (x - \mu)^2)},$$

  and

$$F(x) = \frac{1}{2} + \frac{1}{\pi} arctan(x - \mu), \;\; F^{-1}(t) = tan\left(\pi\left(t - \frac{1}{2}\right)\right) + \mu,$$

  these are both continuous in $\boldsymbol{R}$. Thus the variance of sample median is

$$avar(m_n) = \frac{1/4}{f(F^{-1}(1/2))^2} = \frac{1/4}{f(\mu)^2} = \frac{\pi^2}{4},$$

  and on the other hand

$$avar(\bar{X}_n) = nVar(\bar{X}_n) = Var(_i) = \infty.$$

Example 3.3. *(**Huber loss function**) We consider continuous, symmetric distributions. Let $X \sim F$ be a random variable following a continuous cdf F, with continuous pdf $f = F'$, and suppose we observe n independent copies of $X - \mu$. Suppose also that f is symmetric about 0, and thus $f(x-\mu)$ is symmetric about μ. Consider the following form of the Huber loss*

$$\rho(x) = \begin{cases} k|x| - \frac{1}{2}k^2, & |x| > k \\ \frac{x^2}{2}, & |x| \leq k \end{cases}$$

Then, ρ is continuous and differentiable everywhere, with derivative

$$\rho'(x) = \begin{cases} ksign(x), & |x| > k \\ x, & |x| < k \end{cases}$$

Let $\psi(x) = \rho'(x)$. Notice that the Huber estimator has the correct mean:

$$\begin{aligned} \boldsymbol{E}\psi(X-\mu) &= \int_{\mu-k}^{\mu+k} (x-\mu)f(x-\mu)dx + \int_{-\infty}^{\mu-k} ksign(x-\mu)f(x-\mu)dx + \int_{\mu+k}^{\infty} ksign(x-\mu)f(x-\mu)dx \\ &= \int_{-k}^{k} uf(u)du + \int_{-\infty}^{-k} ksign(u)f(u)du + \int_{k}^{\infty} ksign(u)f(u)du = 0 \end{aligned}$$

Thus, Huber function has the correct asymptotic mean. If $X - \mu$ has pdf f, then μ is the stationary point of $\boldsymbol{E}(\rho(X-b))$. The Huber estimator for the sample is the solution of

$$\hat{\mu} = argmin_{b \in \boldsymbol{R}} \sum_{i=1}^{n} \rho(X_i - b)$$

This can be found by solving $\sum_{i=1}^{n} \psi(X_i - b) = 0$. We wish to find the asymptotic variance of this estimator. To this end, we can do a Taylor expansion of $\sum_{i=1}^{n} \psi(X_i - b)$ at $b = \mu$ to find

$$0 = \sum_{i=1}^{n} \psi(X_i - \hat{\mu}) = \sum_{i=1}^{n} \psi(X_i - \mu) + (\hat{\mu} - \mu) \sum_{i=1}^{n} \psi'(X_i - \mu) + \dots$$

Rearranging,

$$\begin{aligned} \sqrt{n}(\hat{\mu} - \mu) &= \sqrt{n} \frac{\sum \psi(X_i - \mu)}{-\sum \psi'(X_i - \mu)} + \dots \\ &= \frac{-\frac{1}{\sqrt{n}} \sum \psi(X_i - \mu)}{\frac{1}{n} \sum \psi'(X_i - \mu)} + \dots \end{aligned}$$

Where the higher order terms go to zero. By CLT,

$$\sqrt{n}\left(-\frac{1}{n}\sum_{i=1}^{n} \psi(X_i - \mu)\right) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \boldsymbol{E}_\mu \psi^2(X-\mu))$$

Therefore,

$$\sqrt{n}(\hat{\mu} - \mu) \xrightarrow[n\to\infty]{P} \frac{-\frac{1}{\sqrt{n}} \sum \psi(X_i - \mu)}{\frac{1}{n} \sum \psi'(X_i - \mu)} \xrightarrow[n\to\infty]{d} \mathcal{N}(0, \frac{\boldsymbol{E}_\mu \psi^2(X-\mu))}{[\boldsymbol{E}\psi'(X-\mu)]^2})$$

For Huber loss we have,

$$\boldsymbol{E}_\mu \psi'(X-\mu) = \int_{\mu-k}^{\mu+k} f(x-\mu)dx = P(|X| \leq k)$$

and

$$\begin{aligned} \boldsymbol{E}_\mu \psi^2(X-\mu) &= \int_{\mu-k}^{\mu+k} x^2 f(x-\mu)dx + k^2 \int_{-\infty}^{\mu-k} f(x-\mu)dx + k^2 \int_{\mu+k}^{\infty} f(x-\mu)dx \\ &= \int_{\mu-k}^{\mu+k} x^2 f(x-\mu)dx + 2k^2 P(X < -k). \end{aligned}$$

We now, calculate it for the Cauchy distribution. We have $\boldsymbol{E}\psi'(X-\mu) = \int_{-k}^{k} f(x)dx = \int_{-k}^{k} \frac{1}{\pi(1+x^2)}dx = \frac{1}{\pi}(\arctan(k) - \arctan(-k))$ *and* $\boldsymbol{E}\psi^2(X-\mu) = \int_{\mu-k}^{\mu+k}(x-\mu)^2 f(x-\mu)dx + \int_{-\infty}^{\mu-k} k^2 f(x-\mu)dx + \int_{\mu+k}^{\infty} k^2 f(x-\mu)dx = \int_{-k}^{k} x^2 \frac{1}{\pi(1+x^2)}dx + \int_{-\infty}^{-k} k^2 \frac{1}{\pi(1+x^2)}dx + \int_{k}^{\infty} k^2 \frac{1}{\pi(1+x^2)}dx$. *The first integral is solved as* $\int x^2 \frac{1}{\pi(1+x^2)}dx = \frac{1}{\pi}(x - \arctan(x))$ *by* $x = tan(u)$ *substitution. Thus we have* $\boldsymbol{E}\psi^2(X-\mu) = \frac{k^2}{2} + \frac{2k}{\pi} - \frac{2}{\pi}\arctan(k)$. *Putting this all together, the asymptotic variance of the Huber estimator is* $\frac{\boldsymbol{E}\psi^2(X-\mu)}{[\boldsymbol{E}\psi'(X-\mu)]^2} = \frac{\pi}{2}\frac{k-k^2\arctan(k)}{\arctan(k)^2}$. *As* $k \to 0$, *this converges to* 1. □

4 Hypothesis testing

This is the engine of data driven science.

4.1 Parametric Hypothesis testing

Consider a sample $X_1, \dots, X_n$ of iid random variables and a statistical model $(E, (\boldsymbol{P}_\theta)_{\theta\in\Theta})$. Let Θ_0 and Θ_1 be a partition of Θ, not necessarily exhaustive, i.e $\Theta_1 \cup \Theta_2 \subseteq \Theta$. Consider the two hypotheses:

$$H_0 : \theta \in \Theta_0$$
$$H_1 : \theta \in \Theta_1$$

Where H_0 is the null Hypothesis and H_1 is the alternative hypothesis. We say we test H_0 *against* H_1. We want to decide whether to reject H_0, or we fail to reject H_0 (we never accept H_0). For a valid test $\Theta_1 \cap \Theta_2 = \emptyset$ is required, but $\Theta_1 \cup \Theta_2 = \Theta$ is not a requirement. Θ_k is a simple hypothesis if $\Theta_k = \{\theta_k\}$ generally true for null. Θ_k is a composite hypothesis if $\Theta_k = \{\theta : \theta > \theta_k\}$, $\Theta_k = \{\theta : \theta < \theta_k\}$, or $\Theta_k = \{\theta : \theta \neq \theta_k\}$. We can have a two-sided test, e.g. $H_0 : \theta = \theta_0$, $H_1 : \theta \neq \theta_0$, or one-sided test, e.g., $H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$ or $H_0 : \theta \geq \theta_0, H_1 : \theta < \theta_0$.

In A/B testing or two-sample test, e.g. drug trial, we observe two independent samples - test $(X_i \sim \mathcal{N}(\mu_d, \sigma_d^2), i = 1, \dots, n)$ and control group $(X_j \sim \mathcal{N}(\mu_c, \sigma_c^2), j = 1, \dots, m)$ where the Hypothesis testing problem could be

$$H_0 : \mu_d \leq \mu_c$$
$$H_1 : \mu_d > \mu_c$$

e.g., $\mu_d > 0$ is the expected positive effect of a drug for a patient that has used the drug, and μ_c is the observation for a patient who has used the placebo. The data is only used to try to reject H_0, i.e. H_0 and H_1 do not play a symmetric role. We generally take status quo as H_0 while a potential discovery is treated as H_1. In particular, lack of evidence, does not mean that H_0 is true, but rather we fail to reject H_0.

A **test** is a statistic $\psi \in \{0, 1\}$ that does not depend on the unknown quantities and such that - if $\psi = 0, H_0$ is not rejected; and if $\psi = 1, H_0$ is rejected. This can be written as an indicator function $\psi = \mathbf{1}_R$, where R is the rejection region. A test can make two types of errors -

- **type 1 error** where H_0 is true but we reject the Null, and
- **type 2 error** in which H_1 is true and we fail to reject the Null.

Both can be computed from the **power function** $\beta(\theta) = \boldsymbol{P}[\psi = 1]$. The trade-off between the two is shown here:

- If $\theta \in \Theta_0$ (H_0 is true): then $\beta(\theta) = \boldsymbol{P}_\theta[\psi$ makes an error of type 1]. Here we want $\beta(\theta)$ to be small.
- If $\theta \in \Theta_1$ (H_1 is true): then $\beta(\theta) = 1 - \boldsymbol{P}_\theta[\psi$ makes an error of type 2]. Here we want $\beta(\theta)$ to be large.

The **Neyman-Pearson paradigm** is followed, which is a multi-objective problem with Type 1 error deemed more important of the two: Make sure that $\boldsymbol{P}$[Type 1 error] $\leq \alpha \in (0, 1)$ and then minimize $\boldsymbol{P}$[Type 2 error] subject to this constraint. This α is called **level** of the test. Here we will simply focus on the first, without minimizing the second. To identify $\theta \in \Theta_0$ to compute $\boldsymbol{P}[\psi = 0]$, we use $\boldsymbol{P}_\theta[\psi = 1] \leq \alpha, \forall\theta \in \Theta_0 \iff \max_{\theta\in\Theta_0} \boldsymbol{P}[\psi = 1] \leq \alpha$. For tractability, we employ CLT and test $\psi = \psi_n$ and use asymptotic level α if

$$\lim_{n\to\infty} \max_{\theta\in\Theta_0} [\psi_n = 1] \leq \alpha.$$

The **power of a test** is defined as $\boldsymbol{P}_{\theta\in\Theta_1}[\psi = 1]$. There is a duality between building a test and confidence interval. Say we have $\mathcal{I} = [A, B]$ as the CI at level $1 - \alpha$ for a parameter θ, i.e. $\boldsymbol{P}_\theta[\theta \in [A, B]] \geq 1 - \alpha$. We want to use this $\mathcal{I}$ to build a test at level α for $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$. A natural candidate for the test then is $\psi = \mathbf{1}_{\theta_0\notin\mathcal{I}}$. The level of the test is $\boldsymbol{P}_{\theta_0}[\psi = 1] = \boldsymbol{P}_{\theta_0}[\theta_0 \notin \mathcal{I}] = 1 - \boldsymbol{P}_{\theta_0}[\theta_0 \in \mathcal{I}] \leq 1 - \alpha$. Thus, ψ is a test with level $1 - \alpha$. This means that if we repeat the experiment many times, at most α of the tests will make an error of type 1. The **p-value** of a test ψ is the smallest level α at which ψ rejects H_0, i.e. p-value $\leq \alpha \iff H_0$ is rejected by ψ, at the level α. p-value of 5% is strong, 1% is very strong and 0.1% is indisputable.

Example 4.1. *Let $X_1, \dots, X_n \overset{i.i.d.}{\sim} Poiss(\lambda)$, for some unknown $\lambda > 0$ and let λ_0 be a fixed known positive number. The type 1 error is $\boldsymbol{P}_{\lambda\in\Theta_0}[\psi(X_n) = 1]$ and similarly we have type 2 error $\boldsymbol{P}_{\lambda\in\Theta_1}[\psi(X_n) = 0]$. The level of test is α, such that $\sup_{\lambda\in\Theta_0} \boldsymbol{P}[\psi(X_n) = 1] \le \alpha$. To make it tractable we look at asymptotic level as $n \to \infty$. Consider the following hypotheses and give a test of the form $\mathbf{1}_{T_n>s}$ with asymptotic level 5%.*

- *$H_0 : \lambda = 2, H_1 : \lambda \neq 2$. This is a two sided test. We can take the estimator as $\hat{\lambda} = \bar{X}_n$. By LLN, $\hat{\lambda} \xrightarrow[n\to\infty]{P_\lambda} \lambda$ and by CLT $\sqrt{n}\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$. Thus, we can take $T_n = \sqrt{n}\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda}}$, with $\psi = \mathbf{1}_{T_n>s}$.*
  - *To control type 1 error we consider $\lambda \in \Theta_0 \implies \lambda = 2$ and take the statistic as $T_n = |\sqrt{n}\frac{\hat{\lambda}-2}{\sqrt{2}}|$. Thus, we have the type 1 error as $\boldsymbol{P}_2[\psi(X_n) = 1] = \boldsymbol{P}_2[T_n > s] = \boldsymbol{P}_2[|\sqrt{n}\frac{\hat{\lambda}-2}{\sqrt{2}}| > s] \xrightarrow[n\to\infty]{(d)} \boldsymbol{P}[|Z| > s] = 2(1 - \Phi(s)) = \alpha \implies s = q_{\alpha/2}$, which is the $1 - \frac{\alpha}{2}$ quantile of a normal distribution.*
  - *To analyze asymptotic type 2 error we consider the case $\lambda \neq 2$. Under this $\boldsymbol{P}_\lambda[T_n \le s] = \boldsymbol{P}_\lambda[\sqrt{n}|\frac{\hat{\lambda}-2}{\sqrt{2}}| \le s] \xrightarrow[n\to\infty]{P} 0$, because by LLN $|\frac{\hat{\lambda}-2}{\sqrt{2}}| \xrightarrow[n\to\infty]{P_\lambda} \neq 0$.*
  - *Thus the power function $\beta(\lambda) = \boldsymbol{P}_{\lambda\neq2}[\psi(X_n) = 1] = \boldsymbol{P}_{\lambda\neq2}[T_n > s]$ peaks to a value of $1 - \alpha$ at $\lambda = 2$ and drops of both sides.*
- *$H_0 : \lambda \le 2, H_1 : \lambda > 2$. This is a one sided test. We write CLT as $\sqrt{n}\frac{\lambda-\hat{\lambda}}{\sqrt{\lambda}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$ and we define $T_n = \sqrt{n}\frac{2-\hat{\lambda}}{\sqrt{2}}$.*
  - *If $\lambda > 2$, type 1 error $\boldsymbol{P}_\lambda[T_n > s] = \boldsymbol{P}_\lambda[\sqrt{n}\frac{2-\hat{\lambda}}{\sqrt{2}} > z]$. Now, $\frac{2-\hat{\lambda}}{\sqrt{2}} \to \frac{2-\lambda}{\sqrt{2}} < 0$. Thus asymptotically, type 1 error goes to 0.*
  - *For the case of $\lambda = 2$, type 1 error $\boldsymbol{P}_2[Tn > s] = \boldsymbol{P}_2[n\frac{2-\hat{\lambda}}{\sqrt{2}} > s] \xrightarrow[n\to\infty]{d} \boldsymbol{P}[Z > s] = 1 - \Phi(s) = \alpha \iff s = q_\alpha$.*
  - *Finally, for $\lambda < 2$ then we can show similarly, that the type 2 error goes to zero as $n \to \infty$.*
- *$H_0 : |\lambda - 2| \le 1, H_1 : |\lambda - 2| > 1$. This is a composite test. We define the test as $\psi = \mathbf{1}_{\{T_n^l>s_l \text{ or } T_n^t>s_r\}}$ with $T_n^l = \sqrt{n}\frac{1-\hat{\lambda}}{\sqrt{1}} = \sqrt{n}(1 - \hat{\lambda})$ and $T_n^r = \sqrt{n}\frac{\hat{\lambda}-3}{\sqrt{3}}$. For $\lambda = 1$ we have type 1 error as $\boldsymbol{P}_1[T_n^l > s_l \text{ or } T_n^r > s_r] \le \boldsymbol{P}_1[T_n^l > s_l] + \boldsymbol{P}_1[T_n^r > s_r]$. Now, $\boldsymbol{P}_1[T_n^l > s_l] \to \boldsymbol{P}[Z > s_l] = \alpha$ and $\boldsymbol{P}_1[T_n^r > s_r] \to 0$ by LLN. The other case works similarly. For $\lambda = 3$ we have $\boldsymbol{P}_3[T_n^r > s_r] \to \boldsymbol{P}[Z > s_r] = \alpha$. For the case of $\lambda \in (1, 3)$ we get vanishing type 1 error.*

Example 4.2. *We are given $X_1, \dots, X_n \sim Ber(p_x)$ and $Y_1, \dots, Y_n \sim Ber(p_y)$ as iid two set of samples which are independent. We want to test $H_0 : p_x = p_y, H_1 : p_x \neq p_y$. We want to find a test statistic $\psi = \mathbf{1}_{T_n>s}$ at level α.*

To consider $\hat{p}_x - \hat{p}_y$ we need to use multidimensional CLT with delta method. $\hat{p}_x - \hat{p}_y = g(\hat{p}_x, \hat{p}_y)$, with $g(x, y) = x - y$. This gives

$$\sqrt{n}\left(\begin{bmatrix}\hat{p}_x\\ \hat{p}_y\end{bmatrix} - \begin{bmatrix}p_x\\ p_y\end{bmatrix}\right) \xrightarrow[n\to\infty]{(d)} \mathcal{N}\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}p_x(1-p_x) & 0\\ 0 & p_y(1-p_y)\end{bmatrix}\right)$$

Applying delta method with $g(x, y) = x - y$ we get

$$\sqrt{n}(\hat{p}_x - \hat{p}_y - (p_x - p_y)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, p_x(1 - p_x) + p_y(1 - p_y))$$

This gives the test statistic as $T_n = \sqrt{n}\frac{\hat{p}_x-\hat{p}_y-(p_x-p_y)}{\sqrt{p_x(1-p_x)+p_y(1-p_y)}}$. For H_0 we have $p_x = p_y = p \in (0, 1)$ and we can take the plug in estimate as $\hat{p} = \frac{1}{2}(\hat{p}_x + \hat{p}_y)$. And using Slutsky's theorem we have $\sqrt{n}\frac{\hat{p}_x-\hat{p}_y}{\sqrt{2\hat{p}(1-\hat{p})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$. Thus, the test statistic is $T_n = |\sqrt{n}\frac{\hat{p}_x-\hat{p}_y}{\sqrt{2\hat{p}(1-\hat{p})}}|$ To adjust s for level of test α, we can easily see $s = q_{\alpha/2}$. In the case, the null hypothesis is not true $p_x \neq p_y$, we have $T_n \to +\infty$. Thus we have vanishing type 2 error. □

We can formalize the machinery to **Wald test**, which only guarantees asymptotic level (the alternative to it is **T-test**). For a statistical model we use the asymptotic normality of the estimator $\frac{\hat{\theta}-\theta}{\sqrt{\widehat{Var}[\hat{\theta}]}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$. We have the test statistic $W = \frac{\hat{\theta}-\theta_0}{\sqrt{\widehat{Var}[\hat{\theta}]}}$.

- For $H_0 : \theta = \theta_0, H_1 : \theta \neq \theta_0$ the Wald Test ψ is $\mathbf{1}\{|W| > q_{\alpha/2}\}$ with p-value $\boldsymbol{P}[|W| > |W^{obs}|]$
- for $H_0 : \theta \leq \theta_0, H_1 : \theta > \theta_0$ the Wald Test is $\mathbf{1}\{W > q_\alpha\}$ with p-value $\boldsymbol{P}[W > W^{obs}]$
- for $H_0 : \theta \geq \theta_0, H_1 : \theta < \theta_0$ the Wald Test is $\mathbf{1}\{W < -q_\alpha\}$ with p-value $\boldsymbol{P}[W < W^{obs}]$.

For asymptotic normal **two sample Wald test** we define $\hat{\theta} = \bar{X}_n - \bar{Y}_m$, then we have $\frac{\hat{\theta}-\theta}{\sqrt{Var(\hat{\theta})}} \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0,1)$ but with $Var(\hat{\theta}) = Var(\bar{X}_n) + Var(\bar{Y}_m) = \frac{\sigma_1^2}{n} + \frac{\sigma_2^2}{m}$ which can be estimated as $\hat{\sigma}_1^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$ and $\hat{\sigma}_2^2 = \frac{1}{m}\sum_{i=1}^{m}(Y_i - \bar{Y}_m)^2$, both of which are consistent, so by Slutsky we have $\frac{\hat{\theta}-\theta}{\sqrt{\frac{\hat{\sigma}_1^2}{n}+\frac{\hat{\sigma}_2^2}{m}}} \xrightarrow[n\to\infty,\ m\to\infty]{(d)} \mathcal{N}(0,1)$

Another general approach to constructing test is the **Likelihood ratio test** which is used to decide between two hypotheses of the form $H_0 : \theta = \theta_0, H_1 : \theta = \theta_1$. The test is of the form $\psi_X = \mathbf{1}\{\frac{\mathcal{L}_n(X_n,\theta_1)}{\mathcal{L}_n(X_n,\theta_0)} > C\}$ where C is a threshold to be determined. By Wilks' theorem under H_0 we have

$$T_n = 2(\ell_n(\hat{\theta}_n^{MLE}) - \ell_n(\hat{\theta}_n^c)) \to \chi^2_{d-r},$$

where d is the full dimension of parameter θ and last r coordinates are fixed in the alternative. Notice that $\ell_n(\hat{\theta}_n^c)$ is the maximum likelihood in the space Θ_0.

For d dimensional θ **Multidimensional Wald test** uses

$$T_n = n(\hat{\theta}_n^{MLE} - \theta_0)^T I(\hat{\theta}_n^{MLE})(\hat{\theta}_n^{MLE} - \theta_0) \xrightarrow[n\to\infty]{(d)} \chi^2_d$$

with $\psi = \mathbf{1}_{T_n > q_\alpha}$ where q_α is the $(1-\alpha)$-quantile of χ^2_d.

Finite sample case, T-test: Sometimes the sample size are too small to apply CLT/Slutsky which is central to the Wald test, e.g., early phases of clinical trials. We can tackle this case if we assume that our data is normally distributed. The test statistic $\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}$ is a standard Gaussian, but we still need to deal with σ in finite sample where Slutsky is not valid. We can replace σ with the unbiased estimator $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$. We then look at $\sqrt{n}\frac{\bar{X}_n-\mu}{S_n} = \frac{\sqrt{n}\frac{\bar{X}_n-\mu}{\sigma}}{\sqrt{\frac{S_n^2}{\sigma^2}}}$. This is related to a χ^2 distribution. The variable $Z_1^2 + \cdots + Z_k^2$ is χ^2_k where $Z_1, \ldots, Z_k \overset{iid}{\sim} N(0,1)$. Note that $\boldsymbol{E}[\chi^2_k] = k$ and $Var[\chi^2_k] = 2k$. Cochran's theorem states that for $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ and $S_n^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, then $\frac{(n-1)S_n^2}{\sigma^2} = \sum_{i=1}^{n}\left(\frac{X_i-\bar{X}_n}{\sigma}\right)^2 \sim \chi^2_{n-1}$; and $\bar{X}_n$ and S_n^2 are independent random variables. Therefore, the original statistic $\propto \frac{Z}{\sqrt{\frac{V}{n-1}}}$, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_{n-1}$ are independent. This quantity is pivotal as it does not depend on μ and σ. $\frac{Z}{\sqrt{\frac{V}{k}}}$ is **the Student's T distribution** with k degrees of freedom, where $Z \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_k$ are random variables. by LLN as $k \to \infty$ $\sqrt{V/k} \to \boldsymbol{E}[Z^2] = 1$ and hence the pivot tends to a Gaussian. For $k \geq 30$ t-distribution is almost same as the Gaussian for all practical purposes. We can now define the Student's T test for one sample. Given $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$ with both μ and σ^2 are unknown. We know that $\sqrt{n}\frac{\bar{X}_n-\mu}{S_n} \sim T_{n-1}-$distribution, $\forall\ n$.

- We can construct two-sided test for $H_0 : \mu = \mu_0, H_1 : \mu \neq \mu_0$ by T-Test ψ, $\mathbf{1}\{|T| > q_{\alpha/2}^{t_{n-1}}\}$, where $T = \sqrt{n}\frac{\bar{X}_n-\mu_0}{S_n}$ and $\boldsymbol{P}[t_{n-1} > q_\alpha^{t_{n-1}}] = \alpha$, at non-asymptotic level α. The p-value can be similarly calculated from $\boldsymbol{P}[|T| > |T^{obs}|]$
- For one sided test $H_0 : \mu \leq \mu_0$, $H_1 : \mu > \mu_0$ we have $\psi = \mathbf{1}\{T > 1_\alpha^{t_{n-1}}\}$.
- And for testing $H_0 : \mu \geq \mu_0, H_1 : \mu < \mu_0$ we have the test $= \mathbf{1}\{T < -q_\alpha^{t_{n-1}}\}$.

For a two sample T-test we have the hypothesis $H_0 : \mu_d - \mu_c \leq 0, H_1 : \mu_d - \mu_c > 0$ with two independent samples $X_1, \ldots, X_n \overset{iid}{\sim} \mathcal{N}(\mu_d, \sigma_d^2)$ from the test group, and $Y_1, \ldots, Y_m \overset{iid}{\sim} \mathcal{N}(\mu_c, \sigma_c^2)$ from the control group. We have $T = \frac{\bar{X}_n - \bar{Y}_m - (\mu_d - \mu_c)}{\sqrt{\frac{S_d^2}{n} + \frac{S_c^2}{m}}} \sim t_N$ where $N = \frac{\left(\frac{S_d^2}{n} + \frac{S_c^2}{m}\right)^2}{\frac{S_d^4}{n^2(n-1)} + \frac{S_c^4}{m^2(m-1)}} \geq \min(n, m)$ (Welch-Satterthwaite Formula). If we use $min(n, m)$ we get more conservative p-value, but easy to calculate, if you reject with $min(n, m)$ then we will reject with N as well, but not vice-versa.

4.2 Multiple Hypothesis Testing

When testing one hypothesis, we try to control type 1 error of rejecting the null hypothesis even though it is true, by setting the significance level $\alpha = \boldsymbol{P}[\text{reject } H_0 | H_0 \text{ true}]$ to be a small value. When performing multiple hypothesis tests, setting the significance level for all tests to some fixed α may not be enough to control false significance. If $F(t)$ denotes the proportion of p-values we expect to see that are smaller than t, then by definition of level we can say $F(t) \leq t$. But, since we are working with p-value for t-test we also have that $F(t) = t$ which is the CDF of a $Unif(0, 1)$. For Multiple hypothesis testing we have two solutions. In both cases it is easier to work with p-values $P_i = \boldsymbol{P}_{\mu=0}[|T| > |t_i^{obs}|]$, where t_i^{obs} is the observed value of the test statistic for the i-th test and $T \sim t_{n-1}$-distribution.

- **Control Family Wise Error Rate (FWER)**: Find $C_1, \ldots, C_N$ such that $\boldsymbol{P}_{\mu_i=0}[\bigcup_{i=1}^{N} \{|T_i| > C_i\}] \leq \alpha$. In **the Bonferroni method**, to control FWER, we use the Bonferroni correction. Rather than rejecting each test at level α, we use the much smaller level $\frac{\alpha}{N}$. In other words, reject i-th test iff $p_i < \frac{\alpha}{N}$. Often this is very conservative and results in no discovery. Here, $FWER = \boldsymbol{P}_{\mu_i=0}[\bigcup_{i=1}^{N} \{P_i < \frac{\alpha}{N}\}] = \boldsymbol{P}_{\mu_i=0}[\bigcup_{i=1}^{N} \{|T_i| > q_{\frac{\alpha}{2N}}^{t_{n-1}}\}] \leq \sum_{i=1}^{N} \boldsymbol{P}_{\mu_i=0}[|T_i| > q_{\frac{\alpha}{2N}}^{t_{n-1}}] \leq \sum_{i=1}^{N} \frac{\alpha}{N} = \alpha$.
- **Control False discovery Rate (FDR)**: Find $C_1, \ldots, C_N$ such that

$$FDR = \boldsymbol{E}\left[\frac{\#\{i : |T_i| > C_i \ \& \ \mu_i = 0\}}{\#\{i : |T_i| > C_i\}}\right] = \boldsymbol{E}_{\mu=0}\left[\frac{\#\text{ of False discoveries}}{\#\text{ of discoveries}}\right] \leq \alpha$$

  To control FDR, we use the **Benjamini-Hochberg** method. Intuitively, we should reject tests with the smallest p-values. We order p-values $P_{(1)} < P_{(2)} < \cdots < P_{(N)}$ and call the (i)th test with p-value $P_{(i)}$. We reject all tests (i) such the $i \leq i_{max}$ where $i_{max} := max\{i : P_{(i)} < i\frac{\alpha}{N}\}$. With this procedure $FDR \leq \alpha$. This assumes independence of tests. There are many variations over this procedure, e.g., to account for correlations between p-values.

4.3 Nonparametric Hypothesis testing

Let X be a r.v. given iid copies of X we want to answer the question if X have distribution $\mathcal{N}(0, 1)$. These are goodness of fit tests: we want to know if the hypothesis distribution is a good fit for the data, or is the real distribution outside the hypothesis space. In parametric hypothesis testing, we allow only a small set of alternatives, where as in the goodness of fit testing, we allow the alternative to be anything.

Discrete distribution: For the discrete distribution, let $E = \{a_1, \ldots, a_K\}$ be a finite space and $(\boldsymbol{P}_p)_{p \in \triangle_K}$ be the family of all probability distributions on the simplex $\triangle_K = \{\boldsymbol{p} = (p_1, \ldots, p_K) \in (0, 1)^K : \sum_{j=1}^{K} p_j = 1\}$. For $\boldsymbol{p} \in \triangle_K$ and $X \sim \boldsymbol{P}_p$, we have $\boldsymbol{P}_p[X = a_j] = p_j, j = 1, \ldots, K$. We want to test: $H_0 : \boldsymbol{p} = \boldsymbol{p}^0, H_1 : \boldsymbol{p} \neq \boldsymbol{p}^0$ with asymptotic level $\alpha \in (0, 1)$. We can use likelihood ratio test here. The likelihood of the model is $\mathcal{L}_n(X_1, \ldots, X_n, \boldsymbol{p}) = p_1^{N_1} p_2^{N_2} \ldots p_K^{N_K}$, where $N_j = \#\{i = 1, \ldots, n : X_i = a_j\}$. The log likelihood is $\ell(X_1, \ldots, X_n, \boldsymbol{p}) = \sum_{i=1}^{K} N_i \log p_i$ along with the constraint $\sum_{i=1}^{K} p_i = 1$. Writing the Lagrangian and taking derivatives we get the MLE as $\hat{p}_i = \frac{N_i}{N}$, where $N = \sum_{i=1}^{K} N_i$.

If H_0 is true, then $\sqrt{n}(\hat{\boldsymbol{p}} - \boldsymbol{p}^0)$ is asymptotically normal, and the following holds.

$$T_n = n \sum_{j=1}^{K} \frac{(\hat{\boldsymbol{p}}_j - \boldsymbol{p}_j^0)^2}{\boldsymbol{p}_j^0} \xrightarrow[n\to\infty]{(d)} \chi^2_{K-1}.$$

The decrease in degree of freedom by one is because of the fact that $\hat{\boldsymbol{p}} - \boldsymbol{p}^0 \perp \mathbf{1}$. For asymptotic level α we have the test $\psi_\alpha = \mathbf{1}\{Tn > q_\alpha\}$, where q_α is the $(1-\alpha)$-quantile of χ^2_{K-1}. The Asymptotic p-value of this test is $\boldsymbol{P}[Z > T_n | T_n]$, where $Z \sim \chi^2_{K-1}$ and $Z \perp T_n$. More generally, to test if a distribution $\boldsymbol{P}$ is described by some member of a family of discrete distribution $\{\boldsymbol{P}_\theta\}_{\theta \in \Theta \subset \boldsymbol{R}^d}$ where $\Theta \subset \boldsymbol{R}^d$ is d-dimensional, with support $\{0, 1, \ldots, K\}$ and pmf f_θ then

$$T_n : \sum_{j=0}^{K} \frac{\left(\frac{N_j}{n} - f_{\hat{\theta}}(j)\right)^2}{f_{\hat{\theta}}(j)} \xrightarrow[n\to\infty]{(d)} \chi^2_{(K+1)-1-d}.$$

We remove the extra d from the degree of freedom as we have parametrized the probability simplex using d parameters, rest follows from the previous Wald's test formulation of goodness of fit test.

Continuous distribution: If the distribution is continuous with a probability density function one could discretize and use the previous χ^2 goodness of fit test. However one has to decide on number of bins, and bin size. Thus we tackle it directly instead. Like the CDF $F(t) = \boldsymbol{P}[X \le t] = \boldsymbol{E}[\mathbf{1}_{X \le t}]$, we can define the **empirical CDF** as $F_n(t) = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}_{X_i \le t}, \forall t \in \boldsymbol{R}$. We would want $F(t)$, which is non random, to be close to $F_n(t)$, which is a random function since it depends on X_i.

A sequence of functions $g_n(x)$ converges pointwise to a function $g(x)$ if for each x, $\lim_{n\to\infty} g_n(x) = g(x)$. For example in region $x > 1$ $g_n(x) = \frac{1}{x^n}$ converges pointwise to $g(x) = 0$. For any fixed $x > 1$, $\frac{1}{x^n} \xrightarrow[n\to\infty]{} 0$. A sequence $g_n(x)$ converges uniformly to a function $g(x)$ if $\lim_{n\to\infty} \sup_{x \in \boldsymbol{R}} |g_n(x) - g(x)| = 0$. That is, for every $M > 0$, there exists an n_M such that $\sup_x |g_n(x) - g(x)| < M, \forall n \ge n_M$. For example, in the region $x > 2$, $g_n(x) = \frac{1}{x^n}$ converges uniformly to $g(x) = 0$, since $\sup_{x>2} g_n(x) = \sup_{x>2} \frac{1}{x^n} = \frac{1}{2^n} \xrightarrow[n\to\infty]{} 0$. However, the sequence of functions $g_n(x) = \frac{1}{x^n}$ does not converge uniformly to $g(x) = 0$ in the region $x > 1$, since $\sup_{x>1} g_n(x) = \sup_{x>1} \frac{1}{x^n} = 1$, which does not converge to 0 as $n \to \infty$.

By LLN $\forall t \in \boldsymbol{R}$ we have $F_n(t) \xrightarrow[n\to\infty]{a.s.} F(t)$. This only implies pointwise convergence, but not necessarily uniform convergence. However, by Glivenko-Cantelli theorem called **Fundamental theorem of statistics** we have $\sup_{t \in \boldsymbol{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{a.s.} 0$. Thus, $F_n(t)$ uniformly converges to $F(t)$. By CLT, $\forall t \in \boldsymbol{R}$, we have $\sqrt{n}(F_n(t) - F(t)) \xrightarrow[n\to\infty]{(d)} \mathcal{N}(0, F(t)(1 - F(t)))$. This again is pointwise convergence. Using the functional extension of CLT called **Donsker's theorem** we get the uniform asymptotic normality convergence. If F is continuous, then $\sqrt{n} \sup_{t \in \boldsymbol{R}} |F_n(t) - F(t)| \xrightarrow[n\to\infty]{(d)} \sup_{0 \le t \le 1} |\boldsymbol{B}(t)|$, where $\boldsymbol{B}$ is a Brownian bridge on $[0, 1]$, which is a random curve. It is a pivotal distribution and does not depend on the unknown distribution of data, and its quantile can be looked up in a table.

Let $X_1, \ldots, X_n$ be iid real random variables with unknown cdf F and let F^0 be a continuous cdf. We consider the hypothesis $H_0 : F = F^0, H_1 : F \ne F^0$. Let F_n be empirical cdf of the sample, then if $F = F^0$, then $F_n(t) \approx F^0(t), \forall t \in \boldsymbol{R}$. The test statistic for **Kolmogorov-Smirnov test** is $T_n = \sup_{t \in \boldsymbol{R}} \sqrt{n} |F_n(t) - F^0(t)|$. Then by Donsker's theorem, if H_0 is true, then $T_n \xrightarrow[n\to\infty]{(d)} Z$, where Z has a known distribution, supremum of absolute of a Brownian bridge. Thus for a KS test with asymptotic level α we have $\delta_\alpha^{KS} = \mathbf{1}_{T_n > q_\alpha}$, where q_α is the $(1-\alpha)$-quantile of Z, obtained in tables. The p-value of KS test is $\boldsymbol{P}[Z > T_n | T_n]$.

In practice $F^0(t)$ is non-decreasing and $F_n(t)$ is piece-wise constant. Thus, the maximum deviation can only possibly be achieved at the observation points. Let $X_{(i)}$ be the order statistic, then the KS statistic can be

written as

$$T_n = \sqrt{n} \max_{i=1,\dots,n} \left(\max \left(|\frac{i-1}{n} - F^0(X_{(i)})|, |\frac{i}{n} - F^0(X_{(i)})| \right) \right).$$

This is a pivotal statistic, i.e. if H_0 is true, then the distribution of T_n does not depend on the distribution of X_is. Indeed, let $U_i = F^0(X_i), i = 1, \dots, n$ and let G_n be the empirical cdf of $U_1, \dots, U_n$. Then if H_0 is true, then $U_1, \dots, U_n \overset{iid}{\sim} Unif[0,1]$ and $T_n = \sup_{0 \le x \le 1} \sqrt{n}|G_n(x) - x|$. Another critical point to note is that T_n is pivotal for any n, not just for large n. The quantile values can either be read from a table or simulated easily.

There are other goodness of fit tests as well. We want to measure the distance between two function $F_n(t)$ and $F(t)$. There are other ways, leading to other tests: Kolmogorov-Smirnov $d(F_n, F) = \sup_{t \in \boldsymbol{R}} |F_n(t) - F(t)|$, Cramer-Von Mises $d^2(F_n, F) = \int_{\boldsymbol{R}} [F_n(t) - F(t)]^2 dF(t) = \boldsymbol{E}_{X \sim F}[|F_n(X) - F(X)|^2]$, and Anderson-Darling $d^2(F_n, F) = \int_{\boldsymbol{R}} \frac{[F_n(t) - F(t)]^2}{F(t)(1-F(t))} dF(t)$.

Test of Normality: If we want to test if X has Gaussian distribution, with unknown parameters we might want to try the Donsker's theorem. However the theorem is no longer valid! Notice that if true μ, σ^2 are known then by Donker's theorem $T_n = \sup_{t \in \boldsymbol{R}} \sqrt{n}|F_n(t) - \Phi_{\mu,\sigma^2}| \xrightarrow[n \to \infty]{(d)} \sup_{x \in [0,1]} |\boldsymbol{B}(x)|$. However, once we plug in estimators $\hat{\mu}$ and $\hat{\sigma}^2$, and not their true values, then this convergence result no longer holds. However, it is true that under the null hypothesis $H_0 : F = \Phi_{\mu,\sigma^2}$ for some $\mu \in \boldsymbol{R}, \sigma^2 > 0$ the statistic $\tilde{T}_n = \sup_{t \in \boldsymbol{R}} \sqrt{n}|F_n(t) - \Phi_{\hat{\mu},\hat{\sigma^2}}|$ is pivotal. Moreover, the statistic $\tilde{T}_n$ converges in distribution as $n \to \infty$. The quantiles of $\tilde{T}_n$ can be read from tables, and the related test is called **Kolmogorov-Lilliefors test**. This distribution has thinner tails than Kolmogorov-Smirnov test distribution, thus $q_\alpha^{T_n} > q_\alpha^{\tilde{T}_n}$, and you are less likely to reject H_0.

Quantile-Quantile plots: QQ plots provide a visual way to perform GoF tests, giving easy and quick way to see if a distribution is plausible. We want to check visually if the plot of F_n is close to that of F or equivalently if the plot of F_n^{-1} is close to that of F^{-1}. Thus, in the plot we check if the points $\left(F^{-1}(\frac{1}{n}), X_{(1)}\right), \left(F^{-1}(\frac{2}{n}), X_{(2)}\right), \dots, \left(F^{-1}(\frac{n-1}{n}), X_{(n-1)}\right)$ are near the line $y = x$. F_n is not technically invertible but we define $F_n^{-1}(\frac{i}{n}) = X_{(i)}$, the ith largest observation to make it invertible. Some softwares standardize the data and draw, what is called, an AB line instead. Heavy tails can be seen in the QQ plots (sample quantile on y axis) easily, for two sided heavy tails we see point above the line on the right and below the line on the left.

5 Bayesian Statistics

5.1 Prior and Posterior

In a sense, Bayesian inference amounts to having a **likelihood function** $\mathcal{L}_n(\theta)$ that is weighted by prior knowledge on what θ might be. This is useful in many applications. Instead of spitting out $\hat{\theta}$ we will produce the full distribution of the random variable θ. The Bayesian approach is a tool to update our prior belief using the data.

In our statistical experiment $X_1, \ldots, X_n$ are assumed to be iid with Bernoulli r.v. with parameter p conditionally on p. A usual **prior distribution** on Bernoulli p is the $Beta(a, b)$ distribution which has a pdf

$$beta(x; a, b) = \frac{x^{a-1}(1-x)^{b-1}}{\int_0^1 t^{a-1}(1-t)^{b-1}dt},$$

and is defined on the range $[0, 1]$. After observing the available sample $X_1, \ldots, X_n$ we can update our belief about p taking its distribution conditionally on the data. The distribution of p conditionally on the data is called the **posterior distribution**. Here the posterior is again $Beta(a + \sum_{i=1}^{n} X_i, b + n - \sum_{i=1}^{n} X_i)$. When the prior and posterior are of the same family they are called conjugate, and make calculations easy. More generally, if the prior distribution is $\pi(\theta)$ then the posterior is given by

$$\underbrace{\pi(\theta|X_1, \ldots, X_n)}_{posterior} \propto \underbrace{\pi(\theta)}_{prior} \underbrace{\mathcal{L}_n(X_1, \ldots, X_n|\theta)}_{likelihood}$$

5.2 Choosing the prior

Typically, the prior distribution is to be specified in order to take into account previous knowledge about possible values of the parameter. The data dependent likelihood acts as updating our beliefs from prior to posterior. When applying the Bayesian framework, we have considerable freedom in specifying the family of our prior distribution. The following three factors are critical in deciding our prior:

- Whether or not we could specify the parameters of the distribution so that its shape approximates our prior belief.
- Whether or not the support of the distribution is realistic based on our context.
- How tractable it would be to compute the posterior distribution and perform inference from it, given the form of the likelihood function.

How do we pick the prior if we have no prior information about the parameter θ? We could use uniform distribution on finite parameter space but not on infinite space. However, using $\pi(\theta) \propto 1$ can still serve as an uninformative prior, reflecting an equal belief in each possible value of the parameter. This is an **improper** prior, since it is not really a probability distribution. This results in the posterior being the same as the likelihood.

Jeffrey's prior is a non-informative (objective) prior which is invariant under the rescaling of the parameter, and might be improper. This prior depends on the statistical model used for the observation data and the likelihood function.

$$\pi_J(\theta) \propto \sqrt{\det I(\theta)}$$

where $I(\theta)$ is the Fisher information matrix of the statistical model associated with $X_1, \ldots, X_n$. It puts more weight on the points with more information. Jeffrey's prior gives more weight to values of θ whole MLE estimate has less uncertainty. At these high weight points the data gives more information about the parameter, and a small change to θ will influence the data relatively more. Further, on the form of prior, one can show that $I(\theta)(\Delta\theta)^2 = \boldsymbol{E}[(\Delta(\ell(X_i|\theta)))^2] \implies \Delta\theta \propto \frac{1}{\sqrt{I(\theta)}}$ if we hold the expectation constant. This adjustment is based on a quantitative measure of uncertainty and facilitates accurate conversion between parametrizations. The asymptotic standard deviation of the MLE is $I(\theta)^{-\frac{1}{2}}$, to obtain an expression in the same units of the parameter

vector, suggesting the use of square-root in Jeffrey's prior.

Jeffrey's prior satisfies a reparametrization invariance principle: if η is a reparametrization of θ, $\eta = \phi(\theta)$, then we have $\tilde{\pi}(\eta) = \frac{\pi(\phi^{-1}(\eta))}{|\phi'(\phi^{-1}(\eta))|}$ (in higher dimensions we replace ϕ' with $\det J_\phi$, the Jacobian). The pdf satisfies $\tilde{\pi}(\eta) \propto \sqrt{\det \tilde{I}(\eta)}$, where $\tilde{I}(\eta)$ is the Fisher information of the statistical model parametrized by η instead of θ.

5.3 Inference

For $\alpha \in (0,1)$, a Bayesian confidence region with level α is a random subset $\mathcal{R}$ of the parameter space Θ, which depends on the sample $X_1, \ldots, X_n$, such that

$$\boldsymbol{P}[\theta \in \mathcal{R} | X_1, \ldots, X_n] = 1 - \alpha.$$

Note that $\mathcal{R}$ depends on the prior $\pi(.)$. "Bayesian confidence region" and "confidence interval" are two distinct things. In Bayesian case it is the randomness of prior, filtered through to posterior which is being captured by the confidence region. While in the Frequentist case it is the randomness of data $X_1, \ldots, X_n$ which is being captured in the confidence intervals.

We can use the Bayesian framework to estimate the true underlying parameter, since we have the full posterior distribution. In this instance prior is simply acting as a tool to define a new class of estimators. There are many ways to do it. These estimators depend on the choice of prior.

- Posterior mean: $\int_\Theta \theta \pi(\theta | X_1, \ldots, X_n) d\theta$.
- Maximum a posteriori (MAP): $\hat{\theta}^{MAP} = \underset{\theta \in \Theta}{argmax}\ \pi(\theta | X_1, \ldots, X_n)$.

We can essentially do whatever we want with the posterior distribution to accomplish any level of estimation. Bayes estimator, are generally, consistent and asymptotically normal, and in general, this does not depend on the choice of prior. This means that as $n \to n$, the posterior forgets about the prior.

Example 5.1. *Suppose we have the prior on parameter λ given by $\pi(\lambda)$. Conditioned on this λ, we have observations $X_1, X_2, \ldots, X_n \overset{i.i.d.}{\sim} \mathcal{N}(\lambda, 1)$. Compute the posterior distribution $\pi(\lambda | X_1, X_2, \ldots, X_n)$ and then calculate the mean, median, mode for it under the following two priors*

- *Improper prior, $\pi(\lambda) \propto e^{-a\lambda}, \lambda \in \boldsymbol{R}$, and $a \geq 0$:*
- *Proper prior, $\pi(\lambda) = Exp(a) = ae^{-a\lambda}, \lambda \in [0, \infty)$, and $a \geq 0$:*

We notice that the functional form is same for both the prior distribution. The difference is that for improper prior we have $\in \boldsymbol{R}$, while for the proper case we have $\lambda \in [0, \infty)$. For both the cases the general form of the prior is $\pi(\lambda) \propto e^{-a\lambda} \mathbf{1}_{\lambda \in \mathcal{D}}$, where $\mathcal{D}$ is the domain in which λ takes value. The likelihood is given by $\mathcal{L}(X_1, \ldots, X_n | \lambda) \propto exp\left(-\frac{1}{2}\sum_{i=1}^{n}(X_i - \lambda)^2\right)$. Using Bayes' rule we have the posterior as

$$\begin{aligned}
\pi(\lambda | X_1, \ldots, X_n) &\propto \pi(\lambda) \mathcal{L}(X_1, \ldots, X_n | \lambda) \\
&\propto \exp\left(-a\lambda - \sum_{i=1}^{n} \frac{1}{2}(X_i - \lambda)^2\right) \mathbf{1}_{\lambda \in \mathcal{D}} \\
&\propto \exp\left(-\frac{n}{2}\lambda^2 + \left(\sum_{i=1}^{n} X_i - a\right)\lambda\right) \mathbf{1}_{\lambda \in \mathcal{D}} \\
&\propto \exp\left(-\frac{1}{2}\frac{\left(\lambda - \frac{1}{n}\left(\sum_{i=1}^{n} X_i - a\right)\right)^2}{\left(\frac{1}{\sqrt{n}}\right)^2}\right) \mathbf{1}_{\lambda \in \mathcal{D}}
\end{aligned}$$

Now for the case of Improper prior $\mathcal{D} = \boldsymbol{R}$ and we get the posterior $\pi(\lambda | X_1, \ldots, X_n) = \mathcal{N}\left(\frac{1}{n}\left(\sum_{i=1}^{n} X_i - a\right), \frac{1}{n}\right)$. Thus the mean, median, mode are the same in this case, equal to $\frac{1}{n}\left(\sum_{i=1}^{n} X_i - a\right)$. For the case of proper prior

we have $\mathcal{D} = (0, \infty)$ and we are looking at a truncated normal distribution valid only for $\lambda > 0$. Now, in the case $\sum_{i=1}^{n} X_i > a$ the mode is always at 0, otherwise the mode remains at $\frac{1}{n}\left(\sum_{i=1}^{n} X_i - a\right)$. The mean and median now has to be calculated using the truncated distribution. For examples the median can be shown to be $\frac{1}{n}\left(\sum_{i=1}^{n} X_i - a\right) + \frac{1}{\sqrt{n}}\Phi^{-1}\left(1 - \frac{1}{2}\Phi\left(\frac{1}{\sqrt{n}}\left(\sum_{i=1}^{n} X_i - a\right)\right)\right)$. □

Example 5.2. *Assume we have a dataset $X_1, \ldots, X_n$ drawn from a Multinomial$(p_1, \ldots, p_k)$ distribution. The categories are $1, \ldots, k$ and p_j is the probability of the observation from category j. We define N_j as the total number of observations from category j, i.e. $N_j = \#\{X_i = j\}, i = 1, \ldots, n,\ j = 1, \ldots, k$. The joint pdf of this data (conditioned on $(p_1, \ldots, p_k)$) is given by*

$$f(x_1, \ldots, x_n) = \frac{n!}{N_1! \ldots N_k!} p_1^{N_1} \ldots p_k^{N_k},\ \sum_{j=1}^{k} p_j = 1,\ 0 \leq p_j \leq 1.$$

Assume that $(p_1, \ldots, p_k) \sim$ Dirichlet$(\alpha_1, \ldots, \alpha_k)$, where the Dirichlet pdf is given by

$$f(x_1, \ldots, x_k) = \frac{\Gamma(\alpha_1 + \cdots + \alpha_k)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_k)} x_1^{\alpha_1 - 1} \ldots x_k^{\alpha_k - 1},\ \sum_{i=1}^{k} x_i = 1$$

which has a mean $\frac{(\alpha_1, \ldots, \alpha_k)}{\sum_{i=1}^{k} \alpha_i}$. Compute the posterior distribution of $(p_1, \ldots, p_k)$ and it Bayes estimate with MSE as the risk function.

We can calculate the posterior as follows

$$\begin{aligned}
\pi((p_1, \ldots, p_k)|X_1, \ldots, X_n) &\propto \pi(X_1, \ldots, X_n|(p_1, \ldots, p_k))\pi(p_1, \ldots, p_k) \\
&\propto p_1^{N_1} \ldots p_k^{N_k} p_1^{\alpha - 1} \ldots p_k^{\alpha_k - 1} \\
&\propto p_1^{N_1 + \alpha_1 - 1} \ldots p_k^{N_k + \alpha_k - 1}
\end{aligned}$$

This is again a Dirichlet distribution on $(p_1, \ldots, p_k)$ with parameters $(N_1 + \alpha_1 - 1, \ldots, N_k + \alpha_k - 1)$. The Bayes estimator for the MSE risk function is simply the mean of this distribution which is $\frac{(N_1 + \alpha_1, \ldots, N_k + \alpha_k)}{n + \sum_{j=1}^{k} \alpha_j}$. □

6 Linear Models

6.1 Linear Regression

One can simply do linear regression on any data. But that doesn't mean it is the right thing to do. It does not imply causality. However, predicting Y given X is a valid, common reason to use regression. Linear regression proves nothing about the model. Models can always be misspecified. Hopefully they are approximately linear. Leaving out an important covariate will render the model away from truth, however much you try.

We have the data $(X_i, Y_i),\ i = 1, \ldots, n$ are iid from some unknown joint distribution $\boldsymbol{P}$. P can be described entirely by the joint pdf $h(x, y)$. The marginal density of X is $h(x) = \int h(x, y)dy$ and the conditional density $h(y|x) = \frac{h(x,y)}{h(x)}$. This conditional density contains all the information about Y given X. We can also describe the distribution only partially by using the expectation of Y. The **regression function** of Y with respect to X is defined as

$$\boldsymbol{E}[Y|X = x] = \int_{\Omega_Y} yh(y|x)dy$$

which tells us the average value of Y given the knowledge of $X = x$. $\boldsymbol{E}[Y|X]$ is a random variable and has an associated distribution. Notice, that there are other way to summarize the density, like the conditional median $m(x)$, such that $\int_{-\infty}^{m(x)} h(y|x)dy = \frac{1}{2}$. More generally, we could use **conditional quantiles** as the summary (quantile regression). We could also examine conditional variance $\boldsymbol{V}ar[Y|X = x]$, though it does not give any information about the location but only dispersion. We can use M-estimation to estimate the mean, median, and quantiles of Y for a given x even without having to assume a statistical model for the same.

The simplest functional form for the regression function is a constant. The most complex ones are non-parametric with infinite parameters. The simplest non-trivial function is an affine function $f(x) = a + bx$. Under this linearity assumption we get linear regression model. The theoretical linear regression of Y on X is the line $x \mapsto a^* + b^*$x where $(a^*, b^*) = \underset{(a,b)\in\boldsymbol{R}^2}{argmin}\boldsymbol{E}[(Y - a - bX)^2]$. Setting the partial derivatives to zero gives $b^* = \frac{cov(X,Y)}{var(X)}$ and $a^* = \boldsymbol{E}[Y] - b^*\boldsymbol{E}[X]$.

Clearly the points in real life are not exactly on the line $x \mapsto a^* + b^*x$ if $var[Y|X = x] > 0$. The random variable $\varepsilon = Y - (a^* + b^*X)$ is called noise and satisfies $Y = a^* + b^*X + \varepsilon$, with $\boldsymbol{E}[\varepsilon] = 0$ and $cov(X, \varepsilon) = 0$. We now need to estimate a^*, b^* from real data. Assume that we observe n iid random pairs $(X_1, Y_1), \ldots, (X_n, Y_n)$ with same distribution as (X, Y): $Y_i = a^* + b^*X_i + \varepsilon_i$. Using the statistical hammer of replacing expectation by average we get the estimates $(\hat{a}, \hat{b}) = \underset{(\hat{a},\hat{b})\in\boldsymbol{R}^2}{argmin}\frac{1}{n}\sum_{i=1}^{n}(y_i - a_i - bx_i)^2$. Solving this we get $\hat{b} = \frac{\bar{xy}-\bar{x}\bar{y}}{\bar{x^2}-(\bar{x})^2}$ and $\bar{b} = \bar{y} - \hat{b}\bar{x}$. This is called **Least Squares Estimator**. In this particular case, this is precisely what one obtains by taking the least squares solution for the theoretical linear regression problem and replacing each term with their empirical counterparts according to the plug-in-principle. This trick does not always work out in general!

We now generalize this to multidimensional. We note the the column rank and row ranks are always same and a rank-1 matrix can be written as an outer produce and vice-versa. In general, the sum of two matrices can have varying range of ranks, and they can be greater or less than the ranks of matrices that are being summed up. However, $rank(AB) \leq rank(A)rank(B)$. A multivariate regression is given by

$$Y_i = \boldsymbol{X}_i^T\boldsymbol{\beta} + \varepsilon_i,\ i = 1, \ldots, n.$$

where X_i is the explanatory variables or covariates $X_i \in \boldsymbol{R}^p$, and $\beta \in \boldsymbol{R}^p$ is the model parameter. The noise satisfies $cov(\boldsymbol{X}_i, \varepsilon_i) = \boldsymbol{0}$. The **least squares estimate (LSE)** of $\boldsymbol{\beta}^*$ is the minimizer of the sum of squares errors $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}\in\boldsymbol{R}^p}{argmin}\sum_{i=1}^{n}\left(Y_i - \boldsymbol{X}_i^T\boldsymbol{\beta}\right)^2$. We can write this much more concisely in matrix notation as

$$\underset{(n\times 1)}{Y} = \underset{(n\times p)}{X}\underset{(p\times 1)}{\beta} + \underset{(n\times 1)}{\varepsilon}.$$

The LSE now is $\hat{\beta} = \underset{\beta \in \boldsymbol{R}^p}{argmin} \|Y - X\beta\|_1^2$. Linear regression is in fact a parametric statistical model represented by

$$(E, \{\boldsymbol{P}_\beta\}_{\beta \in \Theta}) = \left((\boldsymbol{X}, Y) \in \boldsymbol{R}^d \times \boldsymbol{R}, \begin{Bmatrix} X \sim \mathcal{N}(0, I_d) \\ Y \sim \beta^T X + \varepsilon \\ \varepsilon \sim \mathcal{N}(0,1) \end{Bmatrix}_{\beta \in \boldsymbol{R}^d} \right)$$

This solves to

$$\hat{\beta} = (\boldsymbol{X}^T \boldsymbol{X})^{-1} \boldsymbol{X}^T Y.$$

when $n < p$ we call it the **high dimensional** case. The coefficients are linear combination of Y. $\hat{Y} = \boldsymbol{X}\hat{\beta}$ represents denoised Y. Thus the matrix $\boldsymbol{X}(\boldsymbol{X}^T\boldsymbol{X})^{-1}\boldsymbol{X}^T$ is the projection matrix P, i.e. $PY = \hat{Y}$ (also called the hat matrix) and $P^2 = P$ (idempotent).

To make inference, we need to make more assumptions. Under **deterministic design with Gaussian noise** we assume that the design matrix $\boldsymbol{X}$ is deterministic and $rank(X) = p$. We also assume that the model is **homoscedastic** i.e. ε_i are iid and have a Gaussian distribution $\varepsilon \sim (0, \sigma^2 I_n)$ for some $\sigma^2 > 0$. Under these assumptions we have $Y \sim \mathcal{N}(X\beta, \sigma^2 I_n)$. Under these assumptions, Least Square estimator has the following properties:

- $\hat{\beta}$ is the same as the Maximum Likelihood Estimator.
- Distribution: $\hat{\beta} \sim \mathcal{N}_p(\beta, \sigma^2(X^TX)^{-1})$. The bigger the variance of X lower the variance of $\hat{\beta}$.
- Quadratic risk: $\boldsymbol{E}[\|\hat{\beta} - \beta\|_2^2] = \boldsymbol{E}[\boldsymbol{tr}((\hat{\beta} - \beta)^T(\hat{\beta} - \beta))] = \boldsymbol{tr}\boldsymbol{E}[(\hat{\beta} - \beta)^T(\hat{\beta} - \beta)] = \sigma^2 \boldsymbol{tr}((X^TX)^{-1})$.
- Prediction error: $\boldsymbol{E}[\|Y - X\hat{\beta}\|_2^2] = \boldsymbol{E}[\|(I-P)Y\|_2^2] = \boldsymbol{E}[\|(I-P)Y\|_2^2] + \boldsymbol{E}[\|(I-P)\varepsilon\|_2^2] = \sigma^2(n-p)$. Here we note that $(I - P) \perp Y$ and $(I - P)\varepsilon \sim \mathcal{N}_{n-p}(0, \sigma^2 I_{n-p})$.
- Unbiased estimator of σ^2: $\hat{\sigma}^2 = \frac{1}{n-p}\|Y - X\hat{\beta}\|_2^2$, as suggested by the last property.

We not $(n-p)\frac{\hat{\sigma}^2}{\sigma^2} \sim \chi^2_{n-p}$, and $\hat{\beta} \perp \hat{\sigma}^2$ (Cochran's theorem), because we saw that the two random variables live in orthogonal spaces.

We want to test whether the jth explanatory variable is significant in the linear regression ($1 \leq j \leq p$). The hypothesis is $H_0 : \beta_j = 0 : H_1 : \beta_j \neq 0$. Since we have the full distribution of $\hat{\beta}$, we can easily see that $\hat{\beta}_j \sim \mathcal{N}(\beta_j, \sigma^2\gamma_j)$, where $\gamma_j = [(X^TX)^{-1}]_{ij}$. Standardization gives $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\sigma^2\gamma_j}} \sim \mathcal{N}(0,1)$, however the true value σ^2 are unavailable. So we use the plug in estimate to get $\frac{\hat{\beta}_j - \beta_j}{\sqrt{\hat{\sigma}^2\gamma_j}} \sim t_{n-p}$. For our hypothesis the statistic is $T_n^{(j)} = \frac{\hat{\beta}_j}{\sqrt{\hat{\sigma}^2\gamma_j}}$. We test with non-asymptotic level $\alpha \in (0,1)$ with the rejection region $R_{j,\alpha} = \{|T_n^{(j)}| > q_{\alpha/2}^{t_{n-p}}\}$, where $q_{\alpha/2}^{t_{n-p}}$ is the $(1 - \frac{\alpha}{2})$ quantile of t_{n-p}. We can, similarly, compute the p-values.

An important case occur when p is large and we are testing for their significance simultaneously. That is we want to test whether a group of explanatory variables is significant in the linear regression. $H_0 : \beta_j = 0, \forall j \in S, : H_1 : \exists j \in S, \beta_j \neq 0$, where $S \subseteq \{1, \ldots, p\}$. To account for simultaneous effect we use **Benferroni's test** with the rejection region $R_{S,\alpha} = \bigcup_{i \in S} R_{j,\frac{\alpha}{k}}$, where $k = |S|$. This gives a non-asymptotic level at most α for this test. This is, generally, too conservative resulting in no discovery. False discovery rate is a more prudent concept.

Linear regression exhibits correlation and not causation, because we chose what is X and Y would be. Normality of the noise can be tested by goodness of fit tests on the residual $\hat{\varepsilon}_i$. The deterministic design is not always the most realistic assumption. If X is not deterministic, all the above can be understood conditionally on X, if the noise is assumed to be Gaussian, conditionally on X. This is sometimes called **conditional statistics**.

In the case where homoscedasticity assumption is not valid the error terms ε_i are not iid. We assume that the vector $\varepsilon = [\varepsilon_1, \ldots, \varepsilon_n]^T$ is an n-dimensional Gaussian with mean 0 and known nonsingular covariance matrix Σ. Under these conditions, the MLE is given by minimizing $\hat{\beta} = (Y - X\beta)^T\Sigma^{-1}(Y - X\beta)$ (**weighted least**

squares), which turns out to be

$$\hat{\beta} = (\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1}\boldsymbol{X}^T\Sigma^{-1}Y.$$

To make inference on this we note that $\hat{\beta} \sim \mathcal{N}(\beta, (\boldsymbol{X}^T\Sigma^{-1}\boldsymbol{X})^{-1})$.

6.2 Generalized Linear Models

We want to extend the Gaussian linear models we developed in the last section. The linear model had two components which we are going to extend to get the generalized linear model (GLM):

- Family of distribution: We assumed the conditional distribution $Y|X \sim \mathcal{N}(\mu(X), \sigma^2 I)$ is Gaussian with Y continuous. We can extend it to non continuous and non Gaussian distributions. Under GLM we assume $Y|X \sim$ some distribution from exponential family.
- Regression function: We made the structural assumption of linearity $\boldsymbol{E}[Y|X = x] = \mu(x) = x^T\beta$. Under GLM we introduce a **link function** g, such that $g(\mu(x)) = x^T\beta$ and $\mu(x) = \boldsymbol{E}[Y|X = x]$ is the regression function. We intend to find $f = g^{-1}$ (g is monotone increasing and differentiable) such that $\mu(x) = f(x^T\beta)$. The link function was Identity under Gaussian linear model.

This expressive generalization comes at the cost of more expensive calculation of MLE. Through an appropriate choice of the link function, which depends on the model, we will hope to be able to compute an estimator $\hat{\beta}$, usually the MLE. Consider a model of the number of prey Y that a predator catches given X of preys in the hunting territory. The random component here is, for example, the assumption $Y|X = x \sim Poisson(\mu(x))$ where $\mu(x) = \boldsymbol{E}[Y|X = x]$. The regression function is, for example, $\mu(x) = \frac{mx}{h+x}$ for some unknown $m, h > 0$, where m is the max expected daily preys the predator can cope with and h is the number of preys such that $\mu(h) = \frac{m}{2}$ (half life). $\mu(x)$ is not linear, so we use the **reciprocal link** $g(w) = \frac{1}{w}$. Thus, $g(\mu(x)) = \frac{1}{m} + \frac{h}{m}\frac{1}{x}$, which is linear in $\frac{1}{x}$.

A family of distribution $\{\boldsymbol{P}_\theta : \theta \in \Theta\}, \Theta \subset \boldsymbol{R}^k$ is said to be a **k-parameter exponential family** on $\boldsymbol{R}^q$, if there exist real valued functions $\eta_1, \eta_2, \ldots, \eta_k$ and B of $\boldsymbol{\theta}$; and $T_1, T_2, \ldots, T_k$ and h of $\boldsymbol{y} \in \boldsymbol{R}^q$ such that the density function of $\boldsymbol{P}_\theta$ can be written as

$$f_\theta(y) = h(\boldsymbol{y}) \exp\left(\sum_{i=1}^{k} \eta_i(\boldsymbol{\theta})T_i(\boldsymbol{y}) - B(\boldsymbol{\theta})\right) = h(\boldsymbol{y}) \exp\left(\boldsymbol{\eta}(\boldsymbol{\theta})^T\boldsymbol{T}(\boldsymbol{y}) - B(\boldsymbol{\theta})\right)$$

where $\boldsymbol{\eta}(\boldsymbol{\theta}) = \begin{bmatrix}\eta_1(\boldsymbol{\theta})\\ \vdots \\ \eta_k(\boldsymbol{\theta})\end{bmatrix} : \boldsymbol{R}^k \to \boldsymbol{R}^k$, $\boldsymbol{T}(\boldsymbol{y}) = \begin{bmatrix}T_1(\boldsymbol{y})\\ \vdots \\ T_k(\boldsymbol{y})\end{bmatrix} : \boldsymbol{R}^q \to \boldsymbol{R}^k$, $B(\boldsymbol{\theta}) : \boldsymbol{R}^k \to \boldsymbol{R}$, and $h(\boldsymbol{y}) : \boldsymbol{R}^q \to \boldsymbol{R}$. Normal, Binomial, Poisson, exponential distributions, all belong to the exponential family. Other examples are $Gamma(a, b) : \frac{1}{\Gamma(a)b^a}y^{a-1}e^{-\frac{y}{b}}$ with a the shape parameter, and b the scale parameter (reparametrized by $\mu = ab$ the mean parameter gives $\frac{1}{\Gamma(a)}\left(\frac{a}{\mu}\right)^a y^{a-1}e^{-\frac{ay}{\mu}}$; inverse Gamma$(a, b) = \frac{\beta^\alpha}{\Gamma(\alpha)}y^{-\alpha-1}e^{-\frac{\beta}{y}}$; inverse Gaussian$(\mu, \sigma^2) = \sqrt{\frac{\sigma^2}{2\pi y^3}}e^{-\frac{\sigma^2(y-\mu)^2}{2\mu^2 y}}$; Chi-square, Beta, Negative Binomial distributions. One can write the **canonical form**, for example for $k = 1$, $y \in \boldsymbol{R}$ as $f_\theta(y) = \exp\left(\frac{y\theta - b(\theta)}{\phi} + c(y, \phi)\right)$ for some known functions $b(.)$ and $c(.)$ and ϕ is known and called **dispersion parameter**. The function $b(\theta)$ is known as the **log-partition function**.

Let the log likelihood be defined as $\ell(\theta) = log(f_\theta(Y))$ The mean $\boldsymbol{E}[Y]$ and variance $\boldsymbol{V}ar[Y]$ can be defined using the identities $E\left[\frac{\partial \ell}{\partial \theta}\right] = 0$ and $\boldsymbol{E}\left[\frac{\partial^2 \ell}{\partial \theta^2}\right] + \boldsymbol{E}\left[\left(\frac{\partial \ell}{\partial \theta}\right)^2\right] = 0$. For the canonical form $\ell(\theta) = \frac{y\theta - b(\theta)}{\phi} + c(y, \phi)$. This gives, $\boldsymbol{E}[Y] = b'(\theta)$ and $\boldsymbol{V}ar[Y] = \phi b''(\theta)$.

For Gaussian linear model the link function was identity. For Poisson data since the data is positive we use log link which maps positive real numbers to the entire real line. For Bernoulli/Binomial data we use logit $\log\left(\frac{\mu(X)}{1-\mu(X)}\right)$, probit $\Phi^{-1}(\mu(X))$, or complementary log-log function $\log(-\log(1 - \mu(X)))$, which maps $(0, 1)$ to the entire real line. The function g that links the mean μ to the canonical parameter θ is called **Canonical Link** $g(\mu) = \theta$. Since, $\mu = b'(\theta)$ we have $g = (b')^{-1}$. If $\phi > 0$ the canonical link function is strictly increasing. We can see the following canonical functions for the widely used exponential distributions

	$b(\theta)$	$g(\mu)$
Normal	$\theta^2/2$	μ
Poisson	e^θ	$\log \mu$
Bernoulli	$\log(1+e^\theta)$	$\log \frac{\mu}{1-\mu}$
Gamma	$-\log(-\theta)$	$-\frac{1}{\mu}$

Let $(X_i, Y_i) \in \boldsymbol{R}^p \times \boldsymbol{R}$, $i = 1, \ldots, n$ be independent random pairs such that the conditional distribution of Y_i given $X_i = x_i$ has density in the canonical exponential family: $f_{\theta_i}(y_i) = \exp\left(\frac{y_i\theta_i - b(\theta_i)}{\phi} + c(y_i, \phi)\right)$. We collect $\boldsymbol{Y} = (Y_1, \ldots, Y_n)^T$, $\boldsymbol{X} = (X_1, \ldots, X_n)^T$. Here the mean $\mu_i = \boldsymbol{E}[Y_i|X_i]$ is related to the canonical parameter θ_i via $\mu_i = b'(\theta_i)$, and μ_i depends linearly on the covariates through a link function $g(\mu_i) = X_i^T\beta$. For a link function g, we have $\theta_i = (b')^{-1}(\mu_i) = (b')^{-1}(g^{-1}(X_i^T\beta)) = h(X_i^T\beta)$, where h is defined as $h = (g \circ b')^{-1}$. If g is the canonical link function, h is the identity, i.e. $g = (b')^{-1}$.

The log-likelihood is given by $\ell_n(\boldsymbol{Y}, \boldsymbol{X}, \beta) = \sum_i \frac{Y_i\theta_i - b(\theta_i)}{\phi} = \sum_i \frac{Y_i h(X_i^T\beta) - b(h(X_i^T\beta))}{\phi}$ upto a constant term. For a canonical link function we have $\ell_n(\boldsymbol{Y}, \boldsymbol{X}, \beta) = \sum_i \frac{Y_i X_i^T\beta - b(X_i^T\beta)}{\phi}$. The log-likelihood is strictly concave using the canonical function when $\phi > 0$. As a consequence the maximum likelihood estimator is unique. On the other hand, if another parametrization is used, the likelihood function may not be strictly concave leading to several local maxima. Taking a derivative of $\ell_n(\boldsymbol{Y}, \boldsymbol{X}, \beta)$ with respect to β and setting it to zero gives set of equations that has no closed form solution and are solved using optimization algorithms (Gradient descent, Newton's method). The MLE for Bernoulli Y and the logit link function is called the **logistic regression**.

Let $\ell_n(\beta)$ be the likelihood function as a function of β for a given X, Y. We can use **gradient descent** to find a local minimum of the negative of the log-likelihood function. The gradient descent optimization algorithm, in general, is used to find the local minimum of a given function $f(x)$ around a starting initial point x_0. Let $\ell_{n,1}(\beta) = -\ell_n(\beta)$. We give a starting point β and then repeatedly do $\Delta\beta = -\nabla\ell_{n,1}(\beta)$, choose a step size t and update $\beta := \beta + t\Delta\beta$, until a stopping criterion is satisfied. This is generally of the form $\|\nabla\ell_n(\beta)\| \leq \varepsilon$ for some very small ε. The implementation of this algorithm requires one to compute gradients of the function $\ell_n(\beta)$ at various points. Hence, the computational complexity of gradient descent boils down to the complexity of evaluating the gradient of the function $\ell_n(\beta)$.

The asymptotic normality of the MLE also applies to GLMs, if β satisfies conditions required for asymptotic normality of the ML. We now describe three main GLMs and detail the hypothesis tests we can perform for them, pointing out some peculiarities about them.

- **Logistic Regression**: If the response is binary $Y_i \in \{0, 1\}$ and the covariates $X_i \in \boldsymbol{R}^p$ with (X_i, Y_i) data we can use logistic regression. We model $Y_i \sim Ber(p_i)$, and hence $P[Y_i = y_i|p_i] = p_i^{y_i}(1-p_i)^{(1-y_i)} = exp\left(y_i\eta_i - \log(1+e^{\eta_i})\right)$, where $\eta_i = \log\frac{p_i}{1-p_i}$ is the natural parameter mapping $(0,1) \to \boldsymbol{R}$ which we model as $\eta_i = X_i^T\beta$, i.e. our *imposed choice* here is linear. Thus, $p_i = \frac{e^{\eta_i}}{1+e^{\eta_i}}$ and we have $Y_i|X_i \sim Ber\left(\frac{e^{X_i^T\beta}}{1+e^{X_i^T\beta}}\right)$. The log-likelihood is given by $\ell_n(Y|X, \beta) = \sum_{i=1}^n \left(Y_i X_i^T\beta - log\left(1 + e^{X_i^T\beta}\right)\right)$ and we find MLE by $\hat{\beta}^{MLE} = \underset{\beta}{argmax}\ \ell_n(Y|X, \beta)$. The Information matrix $I(\beta) = -\boldsymbol{E}[\nabla^2\ell_1(\beta(Y_1, X_1))]$ can be estimated as $\hat{I}(\beta) = \frac{1}{n}\sum_{i=1}^n X_{ik}X_{il}\frac{e^{X_i^T\beta}}{\left(1+e^{X_i^T\beta}\right)^2}$. To test the hypothesis $H_0 : \beta_j = \beta_j^0, H_a : \beta_j \neq \beta_j^0$ we can use asymptotic Wald test $\psi = \mathbf{1}\{T_n^W > q_{\alpha(\chi_1^2)}\}$ with the test statistic $T_n^W = n\frac{\left(\hat{\beta}_j - \beta_j^0\right)^2}{\left(\hat{I}(\hat{\beta})^{-1}\right)_{jj}}$. We can use $\hat{\beta}$ instead of β here due to uniform continuity. To investigate the power of this test we look at the single dimensional case and calculate the probability $P_{\beta\neq 0}[\psi = 1]$. This can be investigated by observing that $\sqrt{n\hat{I}(\hat{\beta})}(\hat{\beta}) \sim \mathcal{N}(\sqrt{nI(\beta)}\beta, 1)$ and thus $T_n^W = n\hat{\beta}^2\hat{I}(\hat{\beta}) \sim \chi_1^2(nI(\beta)\beta^2)$. However, $\lim_{\beta\to\infty} nI(\beta)\beta^2 = \lim_{\beta\to\infty}\frac{n\beta^2e^\beta}{2(1+e^\beta)^2} = 0$. Thus, Wald test is not a good test for high values of β. Likelihood ratio test serves well here. We calculate the restricted estimator $\hat{\beta}^c = \underset{\beta_j=\beta_j^0}{argmax}\ \ell_n(Y|X, \beta)$. The test statistic is $T_n^{LR} = 2\left(\ell_n(Y|X, \hat{\beta}^{MLE}) - \ell_n(Y|X, \hat{\beta}^c)\right) \xrightarrow[n\to\infty]{d} \chi_1^2$ with the test $\psi^{LR} = \mathbf{1}\{T_n^{LR} > q_\alpha(\chi_1^2)\}$. This does not have the problem as Wald test. Logistic regression can run

into the problem of **separation** when Y_i can be perfectly recovered by a linear classifier. Applying a prior on β can resolve it.

- **Poisson Regression**: If the data (Y_i, X_i) is count data, we can assume that $Y_i \sim Poiss(\lambda_i)$. Hence, $P[Y_i = y_i|\lambda_i] = \frac{\lambda_i^{y_i} e^{-\lambda_i}}{y_i!} = \exp(y_i \log \lambda_i - \lambda_i - \log y_i!) = \exp(y_i\eta_i - e^{\eta_i} - \log y_i!)$, where the natural parameter $\eta_i = \log \lambda_i$ mapping $\boldsymbol{R}_+ \to \boldsymbol{R}$ which we model as $\eta_i = X_i^T\beta$. Thus, $\lambda_i = e^{\eta_i}$ and we have $Y_i|X_i \sim Poiss\left(e^{X_i^T\beta}\right)$. The log-likelihood (upto a constant) is given by $\ell_n(Y|X,\beta) = \sum_{i=1}^{n}\left(Y_i X_i^T\beta - e^{X_i^T\beta}\right)$. This gives $\nabla_\beta \ell_n(Y|X,\beta) = \sum_{i=1}^{n}(Y_i - e^{X_i^T\beta})X_i^T$, and we can find numerically the maximum likelihood estimator $\hat{\beta}^{MLE} = \underset{\beta}{argmax}\ \ell_n(Y|X,\beta)$. We can find the Information matrix as $\hat{I}(\hat{\beta}) = \sum_{i=1}^{n} e^{X_i^T\beta}X_i^T X_i$. We are using plug in estimator $\hat{\beta}$ in this expression. This is a concave problem and has a maxima. By the property of MLE we have $\sqrt{n}(I(\beta))^{1/2}(\hat{\beta}^{MLE} - \beta) \xrightarrow[n\to\infty]{d} \mathcal{N}(0, I)$. We can find a rectangular or circular confidence region for this. We can also use this estimate $\hat{I}(\hat{\beta})$, to get $n(\hat{\beta}^{MLE} - \beta)^T\hat{I}(\hat{\beta})(\hat{\beta}^{MLE} - \beta) \xrightarrow[n\to\infty]{d} \chi_p^2$ where $\beta \in \boldsymbol{R}^p$ which has an elliptical confidence region. Poisson distribution has the same mean and variance, this might not be true in the data, a problem of overdispersion. We use Gamma GLM in that case which is more flexible in modelling.

- **Gamma Regression**: Here we assume $Y_i \sim Gamma(1, b_i) = \frac{1}{b_i}e^{-y/b_i}$ which has and expectation of b_i and variance of b_i^2 hence resolving the issue of overdispersion. Hence, $P[Y_i = y_i|b_i] = exp\left(-\frac{y_i}{b} + \log\frac{1}{b_i}\right)$, where the natural parameter is $\eta_i = -\frac{1}{b_i}$ which we model as $\eta_i = -X_i^T\beta$. Thus $b_i = \frac{1}{X_i^T\beta}$ and we have $Y_i|X_i \sim Gamma(1, \frac{1}{X_i^T\beta})$. One also use $e^{X_i^T\beta}$ to keep the variables positive in a Gamma distribution.

6.3 Principle Component Regression

Let $\boldsymbol{X} \in \boldsymbol{R}^d$ and $X_1, \ldots, X_n$ be n independent copies of $\boldsymbol{X}$. In high dimensions we need a dimension reduction technique to simplify the data and make it tractable, say for visualization. For a matrix $\mathbb{X} \in \boldsymbol{R}^{n\times d}$, the empirical mean is $\bar{X} = \frac{1}{n}\mathbb{X}^T\mathbf{1}$ and empirical covariance matrix is $S = \frac{1}{n}\mathbb{X}^T\left(I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T\right)\mathbb{X}$. The projection matrix $H = I_n - \frac{1}{n}\mathbf{1}\mathbf{1}^T$ satisfies $H^k x = Hx$ for $x \in \boldsymbol{R}^n$. It is also a projection onto the subspace of vectors perpendicular to the vector $\mathbf{1} \in R^n$. This can be seen from the fact that $Hx = x$ and $Hx.\mathbf{1} = 0$. Finally, the matrix H projects onto the subspace $\{x : \frac{1}{n}\sum_{i=1}^{n} x^i = 0\} \subset \boldsymbol{R}^n$, i.e. it is a set of vectors having coordinate wise average equal to 0.

Any vector $\boldsymbol{v} \in \boldsymbol{R}^d$ can be converted to a unit vector $\boldsymbol{u} = \frac{\boldsymbol{v}}{\|\boldsymbol{v}\|}$. The projection of a vector $\boldsymbol{x} \in \boldsymbol{R}^d$ onto a unit vector $\boldsymbol{u}$ is defined as $proj_{\boldsymbol{u}}(\boldsymbol{x}) := (\boldsymbol{u}^T\boldsymbol{x})\boldsymbol{u}$. The quantity $\boldsymbol{u}^T S\boldsymbol{u}$ is the empirical variance of our dataset in the direction or $\boldsymbol{u}$, i.e. the variance of the projected data onto $\boldsymbol{u}$. We consider the optimization problem $\underset{\|\boldsymbol{u}\|_2^2=1}{argmax}\ \boldsymbol{u}^T S\boldsymbol{u}$. The solution to this problem $\boldsymbol{u}^*$ defines the direction of maximum variance. Any positive semi-definite matrix S has a **spectral decomposition** $S = PDP^T$, where P is an orthogonal matrix and D is a diagonal matrix. A matrix $P \in \boldsymbol{R}^{d\times d}$ is orthogonal if $PP^T = P^TP = I_d$. The diagonal elements of D are the eigenvalues of S and the columns of P are the corresponding eigenvectors of S. S is semi-definite positive iff all its eigenvalues are non-negative.

Let $X_1, \ldots, X_n \in \boldsymbol{R}^d$ denote a data set consisting of iid vectors. Let S denote the empirical covariance matrix of this data set. We can decompose $S = PDP^T$ where $\begin{bmatrix} v_1 & v_2 & \ldots & v_d \end{bmatrix} = P \in \boldsymbol{R}^{d\times d}$ is orthogonal and D is a diagonal matrix with entries $\lambda_1 \geq \lambda_2 \geq \cdots \geq 0$ of S. D is the empirical covariance matrix of the $P^TX_i \in \boldsymbol{R}^n$ for $i = 1, \ldots, n$. In particular λ_i is the empirical variance of the $v_i^TX_i$ and hence measure the spread of the data in the direction v_i. v_1 now serves as the solution the original problem we were interested in: $\underset{\|\boldsymbol{u}\|_2^2=1}{argmax}\ \boldsymbol{u}^T S\boldsymbol{u}$. If we choose the first $k < d$ eigenvectors, we have dimensionally reduced the data. Thus if we choose $P_k = \begin{bmatrix} v_1 & v_2 & \ldots & v_k \end{bmatrix}$ and use the transformation $Y_i = P_k^TX_i$ to get the dimensionally reduced dataset $Y_i \in \boldsymbol{R}^k$. To choose k, once can use a **scree plot** to find the inflection point; or one can use the criterion $\frac{\lambda_1+\cdots+\lambda_k}{\lambda_1+\cdots+\lambda_d} \geq 1 - \alpha$, for some $\alpha \in (0, 1)$ that determines the approximation error, also called the **variance explained**. For data visualization one takes $k = 2$, or 3.

PCA is an algorithm that reduces the dimension of a cloud of points and keeps its covariance structure intact. In practice this algorithm is used for clouds the points that are not necessarily random. In statistics, PCA can be used for estimation for example of the population covariance matrix. If $n >> d$, then the empirical covariance matrix S is a consistent estimator. If $n << d$, then we can use **sparse PCA**. Sometimes it is known beforehand that Σ is almost row rank. We can then run PCA and set $S \approx S'$ with first k eigenvalues retained and rest zeroed out in D. S' will be a better estimator of S under the low-rank assumption. A theoretical analysis of this will lead to an optimal choice of the tuning parameter k (**random matrix theory**).

We can solve $Y = X^T\beta + \varepsilon,\ \varepsilon \sim \mathcal{N}(0, I_n)$ via OLS to get $\hat{\beta}^{OLS} = \underset{\beta \in \boldsymbol{R}^p}{argmin} \|Y - X^T\beta\|_2^2 = (X^TX)^{-1}X^TY$.

$$\hat{\beta}^{OLS} \sim \mathcal{N}(\beta, PD^{-1}P)$$

where $X^TX = PDP^T$. The risk of the estimator is $E[\|\hat{\beta} - \beta\|_2^2] = E[\|\hat{\beta} - \mu\|_2^2] + E[\|\beta - \mu\|_2^2] = var + bias^2$ where $\mu = E[\hat{\beta}]$. If we have $cov[\hat{\beta}] = \Sigma$ then $E[\|\hat{\beta} - \mu\|_2^2] = \boldsymbol{tr}(\Sigma)$. The quadratic risk can be calculated as $E[\|\hat{\beta}^{OLS} - \beta\|_2^2] = \boldsymbol{tr}(PD^{-1}P) = \sum_{i=1}^{p} \frac{1}{\lambda_i}$. We note that the OLS estimator is unbiased. If, however invertibility is at doubt, i.e. $p >> n$ then we need to subscribe to other methods like **ridge regression** at the expense of biasing the estimates. One can use PCA to reduce the set to lower dimension and regress on those linear combinations. **Principle component regression** is given by $\hat{\gamma} = \underset{\gamma \in \boldsymbol{R}^k}{argmin} \|Y - XP_k\gamma\|_2^2 = (P_k^TX^TXP_k)^{-1}P_k^TX^TY$ with $\hat{\beta}^{PCR} = P_k\hat{\gamma} = P_k(P_k^TX^TXP_k)^{-1}P_k^TX^TY$. To choose k optimally we need to know the true β, which is unknown, so generally scree plot might do in practice. We can note that $\hat{\beta}^{PCR} = P_k(P_k^TX^TXP_k)^{-1}P_k^TX^T(X\beta + \varepsilon)$. Using the observation $P_k^TPDP^TP_k = \begin{bmatrix} \lambda_1 & 0 & \dots \\ 0 & \ddots & \dots \\ 0 & 0 & \lambda_k \end{bmatrix} = D_k$, we have $\hat{\beta}^{PCR} = P_kD_k^{-1}P_k^T(X^TX\beta + X^T\varepsilon)$. This suggests, $E[\hat{\beta}^{PCR}] = P_kD_k^{-1}D_k^TPDP^T\beta = P_kP_k^T\beta$. Thus, PCR coefficients are simply the projection of the true β on the subspace of first k eigenvectors. Further, we can note that $cov[\hat{\beta}^{PCR}] = P_kD_k^{-1}P_k^TX^TXP_kD_k^{-1}P_k^T = P_kD_k^{-1}P_k^T$. Thus we have a degenerate normal distribution

$$\hat{\beta}^{PCR} \sim \mathcal{N}(P_kP_k^T\beta, P_kD_k^{-1}P_k^T).$$

The variance can be calculated as $E[\|\hat{\beta}^{PCR} - E[\hat{\beta}^{PCR}]\|_2^2] = \boldsymbol{tr}(P_kD_k^{-1}P_k) = \sum_{i=1}^{k} \frac{1}{\lambda_i}$. For Ridge regression we have $\hat{\beta}^{RIDGE} = \underset{\beta \in \boldsymbol{R}^p}{argmin} \|Y - X\beta\|_2^2 + \tau\|\beta\|_2^2 = (X^TX + \tau I_p)^{-1}X^TY$. We can see $E[\hat{\beta}^{RIDGE}] = P\tilde{D}P^T\beta$ with $D_{ii} = \frac{1}{1+\tau/\lambda_i}$ which shows the shrinkage. We also have $Var[\hat{\beta}^{RIDGE}] = \sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+\tau)^2}$.

	bias	variance
OLS	0	$\sum_{i=1}^{p} \frac{1}{\lambda_i}$
PCR	$(I - P_kP_k^T)\beta$	$\sum_{i=1}^{k} \frac{1}{\lambda_i}$
RIDGE	$(I - P\tilde{D}P^T)\beta$	$\sum_{i=1}^{p} \frac{\lambda_i}{(\lambda_i+\tau)^2}$