Probability

$P(A\cup B)=P(A)+P(B)-P(A\cap B)$
$P(A\cap B)=P(A)+P(B)-P(A\cup B)$
$P(A\cup B)\le P(A)+P(B), P(A)=1-P(A^c)\ge 0$

⚠ **Bayes Theorem**

$P(A\mid B)=\frac{P(A\cap B)}{P(B)}=\frac{P(A,B)}{P(B)}=\frac{P(A)P(B\mid A)}{P(B)}$
$P(A\cap B)=P(B)P(A\mid B)$

⚠ **Law of Total Probability**

$P(B)=P(A_1\cap B)+...+P(A_n\cap B)$
$=P(A_1)P(B\mid A_1)+...+P(A_n)P(B\mid A_n)$

If Independent:
$P(A\mid B)=P(A), P(A\cap B\mid C)=P(A\mid C)P(B\mid C)$

Pairwise independent A,B,C:
$P(A\cap B)=P(A)P(B), P(B\cap C)$
$=P(B)P(C), P(A\cap C)=P(A)P(C)$

Mutually/Fully independent A,B,C:
$P(A\cap B\cap C)=P(A)P(B)P(C)$

Combinatorics:

# of k combos using n things

Using **{A,B,C} (3 choose 2):**
Combination(no order, no repeat):
$\binom{n}{k}=\frac{n!}{k!(n-k)!}=3$: {A,B}, {A,C}, {B,C}

Using **{A,A,B,B,C,C} (6 choose 2):**
Combination(no order, repeats):
$\binom{n+k-1}{k}=\frac{(n+k-1)!}{k!(n-1)!}=6$: {A,A}, {A,B}...

Using **{A,B,C} (3 choose 2):**
Permutation(unique, no repeat):
$P(n,k)=\frac{n!}{(n-k)!}=6$: AB, BA, AC, CA, BC, CB

Permutation(unique, repeats):
$P(n,k)=n^k=9$: AA, AB, AC, BA, BB, ...

Subsets: # subsets for n things:
$2^n=8$: { }, {A}, {A,B}, {A,B,C}...
Idea: Think about what Θ is

Expectations

$E[X]=\sum_x xp_X(x),\ E[g(x)]=\sum_x g(x)p_X(x)=\mu$
$E[X]=\int_{-\infty}^{\infty}xf_X(x)dx,\ E[g(x)]=\int_{-\infty}^{\infty}g(x)f_X(x)dx$
$E\left[X^2\right]=Var(X)+(E[X])^2$

Linearity of Expectation:
$E[aX+bY+c]=aE[X]+bE[Y]+c$
$E[X\pm Y]=E[X]\pm E[Y]$

If Indepedent:
$E[XY]=E[X]E[Y], E\left[(XY)^2\right]=E\left[X^2\right]E\left[Y^2\right]$

Law of Total Expectation
$E[X]=E[E[X\mid Y]]=\sum_y p_Y(y)E[X\mid Y=y]$
$E[X]=\int_{-\infty}^{\infty}f_Y(y)E[X\mid Y=y]dy$

Iterated Expectation
$E[X]=E[E[X\mid Y]]$
$E[XY]=E[E[XY\mid X]]=E[E[XY\mid Y]]$
$E[XY\mid X]=X\cdot E[Y\mid X],\ \ E[XY\mid Y]=Y\cdot E[X\mid Y]$

Conditional & Joint
$E[x\mid Y=y]=\sum_x x\cdot p_{X\mid Y}(x\mid y)$ or for $g(x)$

if X,Y are indepedent:
$f_{X,Y}(x,y)=f(x)\cdot f(y),\ \ f(x)=\frac{f_{X,Y}(x,y)}{f(y)}$
$f_{X\mid Y}(x\mid y)=f(x)$
deterministic r.v. are always indie

Variance

$Var(X)=E\left[(X-E[X])^2\right]=$
$E\left[X^2\right]-(E[X])^2=\sigma^2,\ \sigma=\sqrt{\sigma^2}$
$Var\left(X^2\right)=E\left[\left(X^2\right)^2\right]-\left(E\left[X^2\right]\right)^2$

Law of Total Variance
$Var(X)=E[Var(X\mid Y)]+Var(E[X\mid Y])$
decompose X into $X\mid Y$, Y should be something that influeces X.

If independent:
$Var(X\pm Y)=Var(X)+Var(Y)$

If dependent:
$Var(X_1+..)=\sum_{i=1}^{n}Var(X_i)+\sum_{i,j}Cov(X_i,X_j)$

Scaling of Var(X)
extract the constant and a^2:
$Var(aX+b)=a^2Var(X)$

Sample Variance
$S_n=\frac{1}{n}\sum^{n}\left(X_i-\overline{X_n}\right)^2, E[S_n]=\frac{n-1}{n}\sigma^2$

Unbiased Sample Variance
$\widetilde{S_n}=\frac{n}{n-1}S_n=\frac{1}{n-1}\sum^{n}\left(X_i-\overline{X_n}\right)^2=$
$\frac{1}{n-1}\left(\sum X^2-n\left(\overline{X}\right)^2\right),\ E\left[\widetilde{S_n}\right]=\sigma^2$

Covariance
$Cov(X,Y)=E[(X-E[X])(Y-E[Y])]$
$=E[XY]-E[X]E[Y]$
$Cov(X,Y)=0$ means X,Y indepdent, only for **Normal** distributions

Correlation Coefficient
$\rho_{XY}=\frac{Cov(X,Y)}{\sigma_X\sigma_Y}\in[-1,1]$

Uniform Distribution

Discrete
$p_X(x)=\frac{1}{b-a+1},\ F_X(k)=\frac{\lfloor k\rfloor-a+1}{n}$
$E[X]=\frac{a+b}{2},\ Var(X)=\frac{(b-a+1)^2-1}{12}$
if k is an integer:
$F_X(k)=\frac{k-a+1}{b-a+1}$

Continous
$f_X(x)=\frac{1}{b-a}x\in[a,b],\ F_X(x)=\frac{x-a}{b-a}$
$E[X]=\frac{a+b}{2},\ \ Var(X)=\frac{(b-a)^2}{12}$
$L(b)=\frac{1}{b^n}1\{\max X_i\le b\}$ on $[0,b]$

Meta

step 1: Understand exactly what the question is asking
step 2: break the question into smaller parts. e.g. How to break up complex $P(A)$ into smaller $P(A_1)P(A_2)\cdot...$

MIXED RV

$P(X=Y)=p$ Y~discrete,
$P(X=Z)=1-p$ Z~continous
CDF: $F(x)=p\cdot F_Y(x)+(1-p)\cdot F_Z(x)$
$E[X]=p\cdot E[Y]+(1-p)\cdot E[Z]$
Derive $f_X(x)$ from $F_X(x)$

Statistical Models

$(E,(P_\theta)_{\theta\in\Theta})$
E: sample space $(X_1...)$
P: family of prob measures on E

POISSON PROCESS

Poisson Distribution:
X = # of events in a fixed time interval.
Idea: Random # of k events in t units of time with rate λ, exact timing of k is random. $F(x)$=prob of x events happending in a fixed t time interval?

General PMF: $P(X=k)=\frac{(\lambda t)^k e^{-\lambda t}}{k!}$
$E[X_t]=Var(X_t)=\lambda t,\ I=\frac{1}{\lambda},\ \lambda=\frac{E[X_t]}{t}$

For 1 unit of time $t=1$: $P(X=k)=\frac{\lambda^k e^{-\lambda}}{k!}$
$E[X]=Var(X)=\lambda,\ \ E\left[X^2\right]=\lambda(1+\lambda)$
$I=\frac{1}{\lambda},\ \ L(\lambda)=\frac{\lambda^{\sum X_i}e^{-n\lambda}}{(x_1!...x_n!)}$

Probability of **0** event in **1** unit time$=e^{-\lambda}$;
1 event$=\lambda e^{-\lambda}$, **2** events$=\frac{\lambda^2e^{-\lambda}}{2}$

When dealing with **time** intervals between events, use the $\sim\exp(\lambda)$:

Exponential Distribution:
X = unit time until **next** event in Poisson Proccess, aka the inter-arrival time.
Idea: Random time len between Fixed Events
$f_X(x)=\lambda e^{-\lambda x},\ F_X(x)=1-e^{-\lambda x}$
The CDF is: $P(1^{st}$ event at exactly x time).
$E[X]=\frac{1}{\lambda},\ E\left[X^2\right]=\frac{2}{\lambda^2},\ Var(X)=(E[X])^2=I(\lambda)=\frac{1}{\lambda^2}$
$P(X\ge a)=\int_a^\infty \lambda e^{-\lambda x}dx=e^{-\lambda a},\ \exp(\lambda)=\sum_{k=0}^{\infty}\frac{\lambda^k}{k!}$
$\widehat{\lambda}^{MLE}=\frac{1}{\overline{X_n}}=\frac{n}{\sum X},\ Var\left(\widehat{\lambda}^{MLE}\right)=\frac{\lambda^2}{n}$
$L(\lambda)=\lambda^n e^{-\lambda\sum X_i},\ l(\lambda)=n\ln\lambda-\lambda\sum X_i$

Typically when we use T_1, T_2 etc. this means "inter-arrival time". T_2=time interval from T_1 to T_2 $\sim\exp(\lambda)$

Properties
- In a Poisson Process, events must be indepdent, and λ must be the same.
- Conversely, each interval can be broken up into independent subintervals.
- This Memoryless property means, given a fixed time t, any sub-intervals$\sim$ Unif$[0,t]$

Sum of Poisson Process:
(λ_i must be indepedent)
$X_1\sim Poi(\lambda_1), X_2\sim Poi(\lambda_2)$
$X=X_1+X_2: X\sim Poi(\lambda_1+\lambda_2)$.
for 1 unit of time, you'd expect to see $\lambda_A+\lambda_B$ events, so at anytime, you expect to see $\frac{\lambda_A}{\lambda_A+\lambda_B}$ type λ_A events.
$P(\text{out of n arrivals are total of k type A events})=$
$\binom{n}{k}\left(\frac{\lambda_A}{\lambda_A+\lambda_B}\right)^k\left(\frac{\lambda_A}{\lambda_A+\lambda_B}\right)^{n-k}$
$Erlang(k)=Erlang\left(\frac{k}{2}\right)+Erlang\left(\frac{k}{2}\right)$

Split of Poisson Process:
$Ber(p),\ \lambda_A=\lambda p,\ \lambda_B=\lambda(1-p)$

Order Statistics:

expected of the T^{th} event$=\frac{1}{\lambda}\sum_{i=1}^{k}\frac{1}{n-i+1}$

e.g. 3 engines, $E[T_3]=\frac{1}{3\lambda}+\frac{1}{2\lambda}+\frac{1}{\lambda}$

$f_{T_3}(t)=F_{T_3}(t)$ means the Prob that the 3^{rd} order stat happened before t, which is the prob that $1^{st}, 2^{nd}, 3^{rd}$ event all each happened

Bernoulli Process

Trials must be independent and p constant. $X\in\{0,1\}$

Single Bernoulli Trial::
$P(X=1)=E[X]=p,\ Var(X)=p(1-p)$
$I(p)=\frac{1}{p(1-p)},\ \widehat{p}^{\text{MLE}}=\frac{\text{Sum}}{n}=\overline{X_n}$
$P(X=x)=\begin{cases}p & \text{if } x=1\\ 1-p & \text{if } x=0\end{cases}$, or equivocally:
$P(X=x)=p^x(1-p)^{1-x},\ \ I(p)=\frac{1}{p(1-p)}$
$L(X)=\prod p^{X_i}(1-p)^{1-X_i}=p^{\sum X_i}(1-p)^{n-\sum X_i}$

n Bernoulli Trials:
The P() of k successes in n trials:
$P(X=k)=L(p)=\binom{n}{k}p^k(1-p)^{n-k}$
Sum of k success in n trials:
$X=k=X_1+...+X_n$
$E[X]=np,\ Var(X)=np(1-p)$

Time until 1^{st} **success:** 1^{st} success at trial k means trials $k-1$ all failed, and trial k succeed (Geometric Distro):
$P(X_1=k)=(1-p)^{k-1}p$
$F_X(k)=P(X\le k)=1-(1-p)^k$

Time of k^{th} **success:** this is an extended case of the above (Negative Geometric).
$P(Y_k=n)=\binom{n-1}{k-1}p^k(1-p)^{n-k}$
means P(k^{th} success at n time)

For Geometrics: $E[Y_k]=\frac{k}{p}$, this means "E[time] until k^{th} success", Assume memorylessness: $P(A\mid B)=P(A)$.
$Var(Y_k)=\frac{k(1-p)}{p^2}$

Merging Bernoulli Processes
$Ber(p+q-pq)=P(p)\cup P(q)$= P(either or both) occurs.

Splitting Bernoulli Processes
$A\sim Ber(pq), B\sim Ber(p(1-q))$
These streams are not independent
$\text{Binom}(p)\approx N(np,np(1-p))$:
$P(X=19)=P(18.5\le X\le 19.5)$

Conditional & Joint

$p(x\mid y)=\frac{p(x,y)}{p(y)}=\frac{p(x)p(y\mid x)}{p_Y(y)}$

Fixed values $X=x$ are different from X,
$Y=X+N, N\sim N(0,1)\to f(Y\mid X=x)\sim N(x,1)$

Multiplication Rule
$p(x,y)=p(y)\cdot p(x\mid y)=p(x)\cdot p(y\mid x)$
$p(x,y,z)=p(x)\cdot p(y\mid x)\cdot p(z\mid x,y)$
$p(x,y\mid z)=\frac{p(x,y,z)}{p(z)}$
if X,Y cond. indie of Z:
$p(x,y\mid z)=p(x\mid z)\cdot p(y\mid z)$
For Joint **Normal** X,Y:
$E[X\mid Y]=\mu_X+\rho\frac{\sigma_X}{\sigma_Y}(y-\mu_Y),\ Var(X\mid Y)=\sigma_X^2(1-\rho^2)$

Marginal using Total Probability
$p(x)=\sum_y p(x,y)=\sum_y p(x\mid y)p(y)$
$f(x)=\int f(x,y)dy=\int f(x\mid y)f(y)dy$
Prob of a joint in region A:
$P((X,Y)\in A)=\iint_{(x,y)\in A}f(x,y)dxdy$

Expected Value Rule

Normal Distribution

Normal Distribution$\sim N\left(\mu,\sigma^2\right)$:
$f(x)=\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)\sim N\left(\mu,\sigma^2\right)$
$F(x)=P(X\le x)=\int_{-\infty}^{x}\frac{1}{\sigma\sqrt{2\pi}}\exp\left(-\frac{(t-\mu)^2}{2\sigma^2}\right)dt$
$E[X]=\mu,\ \ E\left[X^2\right]=\mu^2+\sigma^2$
$E\left[X^3\right]=\mu^3+3\mu\sigma^2,\ \ E\left[X^4\right]=\mu^4+6\mu^2\sigma^2+3\sigma^4$
$Var(X)=E\left[(X-\mu)^2\right]=\sigma^2$
$Var\left(X^2\right)=4\mu^2\sigma^2+2\sigma^4,\ \ Var\left(\overline{X}\right)=\frac{\sigma^2}{n}$
$L\left(\mu,\sigma^2\mid X\right)=\frac{1}{(2\pi\sigma^2)^{\frac{n}{2}}}\exp\left(-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2\right)$
$l(L)=-\frac{n}{2}\ln(2\pi)-\frac{n}{2}\ln\left(\sigma^2\right)-\frac{1}{2\sigma^2}\sum(X_i-\mu)^2$
$\widehat{\mu}^{MLE}=\frac{\sum X}{n}=\overline{X},\ \widehat{\sigma^2}^{MLE}=\frac{\sum\left(X-\widehat{\mu}\right)^2}{n}=S_n$

Standard Normal Distribution$\sim N(0,1)$:
$f(x)=\frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right)$
$F(x)=\varphi(x)=\frac{1}{\sqrt{2\pi}}\int_{-\infty}^{x}\exp\left(-\frac{t^2}{2}\right)dt$
$E[X]=0,\ \ E\left[X^2\right]=1,\ \ E\left[X^3\right]=0,\ \ E\left[X^4\right]=3$
$Var(X)=1,\ \ Var\left(X^2\right)=2,\ \ Var\left(\overline{X}\right)=\frac{1}{n}$
$E\left[\overline{X_n}\right]=0,\ \ Var\left(\overline{X_n}\right)=\frac{1}{n}$
$E\left[\frac{\sum X_i^2}{n}\right]=1,\ \ Var\left(\frac{\sum X_i^2}{n}\right)=\frac{2}{n}$
$E\left[\left(\frac{\sum X_i}{n}\right)^2\right]=\frac{1}{n},\ \ Var\left(\left(\frac{\sum X_i}{n}\right)^2\right)=\frac{2}{n^2}$
$L(0,1\mid X)=\frac{1}{(2\pi)^{\frac{n}{2}}}\exp\left(-\frac{1}{2}\sum X_i^2\right)$
$l(0,1\mid X)=-\frac{n}{2}\ln(2\pi)-\frac{1}{2}\sum X_i^2$
$\widehat{\mu}^{MLE}=\frac{\sum X}{n}=\overline{X},\ \widehat{\sigma^2}^{MLE}=\frac{\sum\left(X-\overline{X}\right)^2}{n}=S_n$

⚠ wolfram's σ in $N(\mu,\sigma)$ is $\sqrt{\sigma^2}$
Linear Shift:
$Y=aX+b, X\sim N\left(\mu,\sigma^2\right), Y=N\left(a\mu+b,a^2\sigma^2\right)$
Independent Sum: $Z=N_1+N_2$,
$Z\sim N\left(\mu_X+\mu_Y,\sigma_X^2+\sigma_Y^2\right)$
Independent Diff: $Z=N_1-N_2$,
$Z\sim N\left(\mu_X-\mu_Y,\sigma_X^2+\sigma_Y^2\right)$
Z-Table:
$\Phi(-2)=P(Y\le -2)=1-P(Y\le 2)=1-\Phi(2)$
Std to N(0,1): $\frac{X-\mu}{\sigma},\ \ X=\mu+\sigma Z$

PDF (gives density, not prob)
- Must be non-negative: $f(x)\ge 0$,
-$P(X=x)=0$ for any deterministic value
- Must Integrate to 1:
$\int_{-\infty}^{\infty}f(x)dx=1$ for $-\infty\to\infty$ ranges must be added up, e.g. $\int_{-\infty}^{1}x<1,\int_{1}^{\infty}x>1$, if $\int f(x)\ne 1$ but it's finite, we can * $\frac{1}{\int f(x)}$ to normalize it to $f(x)=1$, but

Differences Squared Measures

$(X-c)^2$: Emphasize bigger differences, makes finding critical points easier.

$E\left[(X-c)^2\right]$: average deviations.

$E\left[\left(X-\hat{\theta}\right)^2\right]$: **MSE** of an estimator, aka

Quadratic Risk: $R\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right) + \left(Bias\left(\hat{\theta}\right)\right)^2$

e.g. MSE of $\mu - \overline{X}$ is

$\text{standard error}^2 : E\left[\left(\mu-\overline{X}\right)^2\right] = \left(\frac{\sigma}{\sqrt{n}}\right)^2 = \frac{\sigma^2}{n}$

Variance & Sample Variance:

$E\left[(X-E[X])^2\right] = Var(X),\ \ E\left[\left(X-\overline{X}\right)^2\right] \to S_n$

LMS: Least Mean Squared Error, aka Bayes Estimator, minimizes posterior, is usually the $\mu = E[X]$, or the $E[X \mid Y]$ of a joint distribution:

$E\left[\left(X-\hat{X}^{LMS}\right)^2\right] = E\left[(X-E[X \mid Y])^2\right] = Var(X \mid Y)$

Test Statistics

Typically measures the differences between an estimator $\hat{\lambda}$ and the hypothesis λ_0: $\hat{\lambda} - \lambda_0$

ESTIMATORS

Asym. Normal if: $\sqrt{n}\left(\hat{\theta}-\theta\right) \to N\left(0,\sigma^2\right)$

Consistent if: $\hat{\theta} \to \theta$ as $n \to \infty$

Bias: $Bias\left(\hat{\theta}\right) = E\left[\hat{\theta}\right] - \theta$

Quadratic Risk: $R(\theta) = E\left[\left|\hat{\theta}-\theta\right|^2\right]$

Unbiased Estimator

$\hat{\theta}$ such that $bias\left(\hat{\theta}\right) = 0$, which means $\hat{\theta} \to \theta$

if a $\hat{\theta}$ is biased, use linearity of expectations to create a new estimator such that $E\left[\tilde{\theta}\right] = c \cdot E\left[\hat{\theta}\right] = \theta$

Continuous Mapping Theorem

Apply continuous $g(x)$ to a sequence of $\{X_i\} \to X$, then the transformed $\{g(X_i)\}$ will also $\to g(X)$. As in, applying $g(x)$ preserves the convergence property.

Note: plug in estimators in operations, e.g. plug-in $\hat{\lambda}^{MLE} = \frac{1}{\overline{X}}$ into $l(\lambda)$

MLE Estimator

Since $L(X \mid \theta)$ means the likelihood of observing X_i assuming θ is true, MLE aim to find a $\hat{\theta}$ that makes the data most probable: $\hat{\theta}^{MLE} = arg\max_{\theta} L(X \mid \theta)$.

Does not incorp prior $p(\theta)$, unlike **MAP**. The MLE calculated for the prior $p(\theta)$ is called a constrained MLE

wolfram: $\frac{d}{d\theta}\ln\left(\prod f_\theta(x)\right) = 0$

Likelihood: all of $L(X \mid \theta)$ happening together: $L(X \mid \theta) = \prod_{i=1}^{n} f(X_i \mid \theta)$.

Properties

If support of P_θ does't depend on θ, $\theta^\star$ is not at boundary, $I(\theta)$ is invertible:

Consistent: As $n\uparrow$, MLE $\to \theta^\star$ to true θ

Asymptotic Normality: For large n,

$\sqrt{n}\left(\hat{\theta}^{MLE}-\theta^\star\right) \to N\left(0, \frac{1}{I(\theta^\star)}\right)$, $Var\left(\hat{\theta}\right) = \frac{1}{I(\theta)}$,

Θ: Param set. Well specified if the true $\theta^\star \in \Theta$

⚠ sample space must not depend on parameter

⚠ sample space must be the support for the distribution.

i.e. $\left([0,\infty), \left\{N\left(\mu,\sigma^2\right)\right\}\right)$ is not valid because the sample space for a N is all R

Identifiability

θ identifiable only if mapping $\theta \in \Theta \to P_\theta$ is injective (injective: $\theta \neq \theta\prime \Rightarrow P_\theta \neq P_{\theta\prime}$)

LLN

if iid: $\lim_{n\to\infty} \overline{X_n} = E[X] = \mu$

CLT

if iid, and n is large:

$\frac{\sqrt{n}}{\sigma}\left(\overline{X_n}-\mu\right) = \frac{\overline{X_n}-\mu}{\frac{\sigma}{\sqrt{n}}} \to N(0,1)$

$\sqrt{n}\left(\overline{X}-\mu\right) \to N\left(0,\sigma^2\right)$: $\overline{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right)$

need to enlarge $\overline{X}-\mu$ by $\sqrt{n}$, otherwise as $n\uparrow$, $\overline{X}-\mu$ will $\to 0$

Critical Values (z-score)

$a = 0.025$, 2-tails: $\frac{a}{2} = 0.0125$ in each tail $q_{\frac{a}{2}} = 2.24$, 1-tail: $q_a = 1.96$

$a = 0.05$, 2-tails: $\frac{a}{2} = 0.025$ in each tail $q_{\frac{a}{2}} = 1.96$, 1-tail: $q_a = 1.645$.

$a = 0.1$, 2-tails: $\frac{a}{2} = 0.05$ in each tail $q_{\frac{a}{2}} = 1.645$, 1-tail: $q_a = 1.28$.

Composition of $P(A \cap B \cap C)$

$P(A^c) \to P(A) \to P(B \mid A) \uparrow P(A \cap B) \to P(C \mid A \cap B) \to P(A \cap B \cap C)$

Parameters

θ: True but unknown population θ

$\theta^\star$: Asym of θ, is the $\lim$ of $\hat{\theta}_n$

$\hat{\theta}$: point estimator of θ

θ_0: the θ that H_0 hypothesizes

$\tilde{\theta}$: alt estimator, estimator of $\hat{\theta}$

θ_n: sequence of parameters

$\overline{\theta}$: mean of θ, or $\overline{\theta}_n$

$\hat{\theta}^{MLE,MAP}$: specific types of $\hat{\theta}$

Bayesian

Treats θ as random variables.

Uses prior knowledge.

Provide posterior distribution

Makes probabilistic statements about θ.

Frequentist

Treats θ as fixed but unknown.

Relies solely on data.

Provides $\hat{\theta}$ and confidence intervals.

Makes probabilistic statements about the data.

QQ Plot

Visually access what distribution plotted dataset follows, by comparing its quantiles to the quantiles of a

prob that $1^{st}, 2^{nd}, 3^{rd}$ event all each happened before time t, which means

$F_{T_3}(t) = P(T_1 \le t, T_2 \le t, T_3 \le t) = \left(1-e^{-\lambda t}\right)^3$,

differentiate this to get $f_{T_3}(t)$

Gamma Distribution

$f(x;\alpha,\beta) = \frac{\beta^\alpha}{\Gamma(\alpha)} \cdot x^{\alpha-1}e^{-\beta x}$, $Var(X) = \frac{\alpha}{\beta^2}$

$\hat{x}^{Bayes} = E[X] = \frac{\alpha}{\beta}, E\left[X^2\right] = \frac{\alpha(\alpha+1)}{\beta^2}$

$\Gamma(\alpha)$ normalizes the distribution to 1. if `$\alpha$` integer: $\Gamma(\alpha) = (\alpha-1)!$, posterior

$\pi(\theta \mid X) \propto \Gamma(\theta;\alpha,\beta)$, $\hat{\theta}^{MAP} = \text{Mode} = \frac{\alpha-1}{\beta}$

$X_1 \sim \Gamma(\alpha_1,\beta), X_2 \sim \Gamma(\alpha_2,\beta), X_1+X_2 \sim \Gamma(\alpha_1+\alpha_2,\beta)$

For use with time-related: until k^{th}, not limited to just next event:

$f(x) = \frac{\lambda^k x^{k-1}e^{-\lambda x}}{\Gamma(k)}$. $E[X] = \frac{k}{\lambda}, Var(X) = \frac{k}{\lambda^2}$

Erlang: Gamma but k must be integer:

$f(x) = \frac{\lambda^k x^{k-1}e^{-\lambda x}}{(k-1)!}$, $E[X] = \frac{k}{\lambda}, Var(X) = \frac{k}{\lambda^2}$

Beta Distribution

Continuous distribution defined on $[0,1]$. Models proportions, used as prior.

$f(x;\alpha,\beta) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{B(\alpha,\beta)}$

$B(\alpha,\beta) = \int_0^1 t^{\alpha-1}(1-t)^{\beta-1}dt = \frac{\Gamma(\alpha)\Gamma(\beta)}{\Gamma(\alpha+\beta)}$

$E[X] = \frac{\alpha}{\alpha+\beta}$, $Var(X) = \frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$,

Mode$= \frac{\alpha-1}{\alpha+\beta-2}$

When α=β=1, Beta is a **Uniform**.

Confidence Interval

An interval that will contain the true parameter with a likelihood of $1-\alpha$, cannot depend on unknown parameter.

$CI = [\text{Estimate} - \text{Critical Value} \cdot \text{Standard Error}, \text{Estimate} + \text{Critical Value} \cdot \text{Standard Error}]$

$P\left(\overline{X_n} - q_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}} \le \mu \le \overline{X_n} + q_{\frac{\alpha}{2}}\frac{\sigma}{\sqrt{n}}\right) = 1-\alpha$ ⚠$\sqrt{\sigma^2} = \sigma$

Standard Errors Forms

$\frac{\sigma}{\sqrt{n}}, \sqrt{\frac{\sigma^2}{n}}, \sqrt{\frac{p(1-p)}{n}}, \sqrt{\frac{\sigma_1^2}{n_1}+\frac{\sigma_2^2}{n_2}}$ for $\overline{X}_1 - \overline{X}_2$

Plug-in Method: use the $\hat{\theta}$ in variance formulas for standard error.

Hypothesis Tests

⚠ fail to reject H_0 is not accepting H_0

θ, X_i, $\hat{\theta}$: Given sample space, we use samples X_i to construct $\hat{\theta}$ for the θ we are interested in inferencing.

H_0, H_1: We hypothesized a value for θ we call θ_0, H_1 could be $\theta_0 \neq \theta$, (2-tailed) or θ_0 >or< θ (1-tail).

T_n: test statistic, usually $\frac{\hat{\theta}-\theta_0}{SE\left(\hat{\theta}\right)}$

$T_n \mid H_0$ ~a known distribution, the goal is to leverage the known properties of distributions to infer H_0, e.g. Z-test, t-test, χ^2 test, Wald's, LRT.

1 tail: $P(X$ >or< $T_n)$ is the tail to the right or left; **2-tails:** $P(|X| > |T_n|)$ for tails on both side.

Expected Value Rule

$E[g(X,Y)] = \sum_x\sum_y g(x,y)p(x,y)$

$E[g(X,Y)] = \int E[g(x,y) \mid Y = y]f(y)dy$

$E[g(X,Y) \mid Y = y] = \int g(x,y)f(x \mid y)dy$

CDF: $F(x,y) = P(X \le x, Y \le y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(s,t)dsdt$

Convolution

Find the Probability of the Sum of variables: $Z = X+Y$

$p_Z(z) = \sum_x p_X(x)p_Y(z-x)$

$f_Z(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx$

Random # * Random Amounts

N: # of events

X_i: Amount of each event

T: Total Amount: $T = \sum_{i=1}^{N} X_i$

$E[T] = E[N]E[X] = E[N] \cdot \mu$

by law of total variance:

$Var(T) = E[Var(T \mid N)] + Var(E[T \mid N])$

$Var(T) = \sigma^2 \cdot E[N] + \mu^2 \cdot Var(N)$

Fisher Information

⚠ use only <u>ONE</u> observation: $\frac{1}{Var(X)}$

$I(\theta) = E\left[\left(\frac{d}{d\theta}\ln f(X;\theta)\right)^2\right] = -E\left[\frac{d^2}{d\theta^2}\ln f(X;\theta)\right]$

$I(\theta) = Var(l\prime(\theta)) = -E[l\prime\prime(\theta)]$, $\sum I_i(\theta) = n \cdot I(\theta)$

Cramer-Rao Bound: $Var\left(\hat{\theta}\right) \ge \frac{1}{I(\theta)}$

Invariance re-para: $I(\varphi) = I(\theta)\left(\frac{d\theta}{d\varphi}\right)^2$

Matrix: $I_{ij}(\theta) = E\left[\frac{d}{d\theta_i}\ln f(X;\theta) \cdot \frac{d}{d\theta_j}\ln f(X;\theta)\right]$

not well-defined if distro support depends on unknown paramemter (e.g. shifted $\exp(\lambda)$)

$\frac{d^2}{d\theta^2}$ of log-likelihood must exist:

$I(\theta) = Var(l\prime(\theta)) = -E[l\prime\prime(\theta)]$

Wald's Test

Accesses $\hat{\theta}^{MLE}$ with H_0, typically for $H_1 : \hat{\theta} \neq \theta_0$. Assumes $\hat{\theta} \sim N()$.

Z-form: $T_n^{Wald} = \frac{\hat{\theta}-\theta_0}{SE\left(\hat{\theta}\right)} = \frac{\hat{\theta}-\theta_0}{\sqrt{\hat{V}ar\left(\hat{\theta}\right)}} \sim N(0,1)$,

reject H_0 when $|T_n| > z_{\frac{a}{2}}$ (2-tailed).

χ_1^2 form:

$T_n^{Wald} = \left(\frac{\hat{\theta}-\theta_0}{SE\left(\hat{\theta}\right)}\right)^2 = \frac{\left(\hat{\theta}-\theta_0\right)^2}{\hat{V}ar\left(\hat{\theta}\right)} \sim \chi_1^2$

re-para $g(\theta_0) =$, wald will not be robust.

Multi-variate:

$n\left(\hat{\theta}^{MLE}-\theta_0\right)^T \cdot \left(\hat{\theta}^{MLE}-\theta_0\right) \cdot I\left(\hat{\theta}^{MLE}\right) \to \chi_d^2$

$T_n = \|\sqrt{n}I(\theta_0)^{\frac{1}{2}}\left(\hat{\theta}^{\text{MLE}}-\theta_0\right)\|^2 \to \chi_d^2$

Likelihood Ratio Test

Purpose: test if data fits model better with θ_0 or without.

H_0: selectively set hypothesized values to part or all parameters of the target

if any part of the $\int$ "doesn't converge", means inifinite and can't be normalized.

- Evaluating at specific points gives the relative density of that point, not the actual probabilties, to get the actual probability in that range: integrate between the bounds:

$P(a \le X \le b) = \int_a^b f(x)dx$, this gives the CDF for that range.

CDF (gives actual Prob)

$\to 0$ as $x \to -\infty$, $\to 1$ as $x \to \infty$.

Gives the Prob of the interval which the CDF was integrated up to

$F_X(x) = 1$ when $x \ge$ upper bound

non-decreasing, right-continuous.

$P(a < X \le b) = F_X(b) - F_X(a)$

Integrate up to x for the CDF of X

$F_X(x) = \int_{-\infty}^{x} f_X(t)dt = P(X \le x)$

Inversely, to get the PDF of X, differentiate $f_X(x) = \frac{d}{dx}F_X(x)$

Evaluating specific value gives the area to the left: $P(X \le x)$, for the area to right: $1-P(X \le x)$, which gives $P(X > x)$. $Y = F(x), F_Y \sim \text{Uni}(0,1)$

Empirical CDF: step fn that estimates population $F(t)$ on each $\frac{1}{n}$ step.

$F_n(t) = \frac{1}{n}\sum \mathbf{1}\{X_i \le t\}$, eCDF is a step fn so discont, jumps with $\frac{1}{n}$ at each X_i, converge to true $F(t)$. Asym Normal:

$\sqrt{n}(F_n(t) - F(t)) \to N(0, F(t)(1-F(t)))$,

$F(t)(1-F(t))$ is the var of binom, due to the indicator.

PMF

Non-negative: $P(X = x) \ge 0$, Must Sum to 1: $\sum_x p(x) = 1$, $p(x) = P(X = x)$

Linear Transform

$Y = aX+b$, $p_Y(y) = p_X\left(\frac{y-b}{a}\right)$,

$f_Y(y) = \frac{1}{|a|}f_X\left(\frac{y-b}{a}\right)$

g is monotonic: $f_Y(y) = f_X(h(y))\left|\frac{dh}{fy}(y)\right|$,

where h is inverse of g

1) find CDF: $F_Y(y) = P(g(x) \le y)$

2) derive CDF for y to find PDF

Misc

$(a-b)^2 = a^2+b^2-2ab$

$(a+b)^2 = a^2+b^2+2ab$

Log

$\ln(mn) = \ln(m)+\ln(n), \ln\left(\frac{m}{n}\right) = \ln(m)-\ln(n)$

$\ln(m^r) = r\ln(m)$, $\ln(1) = 0$, $\ln(x \le 0)$undefined

remove or add e: $e^{\ln(x)} = x$

$e^{a+b} = e^a \cdot e^b, e^{ab} = (e^a)^b, e^{-x} = \frac{1}{e^x}, e^0 = 1$

add $\ln$ to both sides to remove e:

$e^x = 2 \to x = \ln(2)$

add e to both sides to remove $\ln$:

$\ln(x) = y$ is $e^{\ln(x)} = e^y$ then $x = e^y$

Wolfram: complete the square x^2..

Exponent

$(ab)^x = a^xb^x, (a^x)^y = a^{xy}, a^xa^y = a^{x+y}$

$\sigma^2\cdot(g'(x))^2$ if re-parameterized.
MLE minimizes **KL diver**, can be Biased.
$f(x)$ must be continuously differentiable to find critical values. No sharp corner, vertical tangents, or discontinuous functions.

MAP Estimator

MAP: Maximum Posterior Estimator. Incorporates prior $p(\theta)$, unlike **MLE**.
$\hat{\theta}^{MAP} = arg\max \pi(\theta \mid X) = arg\max L_n(X \mid \theta)\pi(\theta)$.
Find critical value(s) via $\frac{d}{dx}\pi(\theta \mid X) = 0$, verify min or max by $\frac{d^2}{dx^2}\pi(\theta \mid X)$ and plug back into λ^{MAP}.
MAP is the mode of $arg\max \pi(\theta \mid X)$, which means $\hat{\theta}^{MAP}$ is one of the values of θ if θ is a discrete set.
Example:
given $g(x) = \lambda e^{-\lambda}x, \lambda \sim \exp(\theta)$, find λ^{MAP}.
Solution 1.
1) Posterior ∝ Likelihood*Prior. We ignore the marginal $p(X)$ because it acts as a constant here so it will not change $arg\max$ of λ.
2) $\pi(\lambda \mid X) \propto L(X \mid \lambda)\pi(\lambda) = \lambda^n e^{-\lambda\sum X}\cdot\left(\theta e^{-\theta\lambda}\right)$
3) differentiate this posterior with respect to the parameter we're trying to estimate, set it to 0 and solve.
4) wolfram: $\frac{d}{d\lambda}\lambda^n e^{-\lambda\sum X}\cdot\left(\theta e^{-\theta\lambda}\right) = 0$, we get $\lambda^{MAP} = \frac{n}{\sum X + \theta}$
5) wolfram: $\frac{d^2}{d\lambda^2}\ln\left(\lambda^n e^{-\lambda x}\cdot\left(\theta e^{-\theta\lambda}\right)\right) = -\frac{n}{\lambda^2}$
6) plug this back into λ^{MAP} we get $-\frac{\left(\sum X + \theta\right)^2}{n}$, since $\left(\sum X + \theta\right)^2$ must >0, and $n > 0$, 2nd derivative test is neg, which means λ^{MAP} is the maximum.

Canonical Exponential Family

General: $f(x \mid \theta) = h(x)\exp(\eta(\theta)\cdot T(x) - A(\theta))$
1 parameter: $f_\theta(x) = \exp\left(\frac{x\theta - b(\theta)}{\varphi} + c(x,\varphi)\right)$
$= \exp\left(x\cdot\frac{\theta}{\varphi} - \frac{b(\theta)}{\varphi}\right)\cdot e^{c(x,\varphi)}$ try to move all θ related terms inside the exp()
$\theta = \eta(\theta)$: main canonical parameter
$x = T(x)$: sufficient stat, carry main random info on x
$e^{c(x,\varphi)} = h(x)$: base measure, indie of θ
$\frac{b(\theta)}{\varphi} = A(\theta)$: log-partition function, normalizes the pdf to 1
φ: dispersion, effects variance (GLM)
Properties:
$E[X] = b'(\theta),\ Var(X) = b''(\theta)\varphi,\ I(\theta) = \frac{nb''(\theta)}{\varphi}$
⚠ these formulas belongs to $\exp(\eta(\theta)\cdot T(x) - A(\theta))$, when using transformed para, must replace the transformed para with the equal θ.
linear transformations also canon.
Canon Link relateds $\mu = E[X]$ to θ: $g(\mu(x)) = \eta = \theta = (b')^{-1}(\mu(x))$, if $\varphi > 0$, canon link is strictly increasing.
log-L: $l(\theta; x) = \sum^{n}\Big[\ln h(X_i) + \frac{T(X_i)\theta - b(\theta)}{}$

quantiles to the quantiles of a known distribution.
Theoretical quantile-axis: the reference distribution to compare to, the points on the axis are its support.
Sample quantile-axis: the distribution from the sample which we are interested in, values on axis are it's support.

lighter tails
lesser ext values
such as uniform
exp(+)

fatter tails
more ext values
such as t-distro
exp(-)

Canonical Exp Family Example:
Form: $f_\theta(x) = h(x)\exp(\eta(\theta)T(x) - A(\theta))$
$f_\lambda(x) = \frac{x^4}{24\lambda^5}\exp\left(-\frac{x}{\lambda}\right) =$
$\exp\left(-\frac{1}{\lambda}\cdot x\right)\cdot\frac{1}{\lambda^5}\cdot\frac{x^4}{24} =$
$\exp\left(-\frac{1}{\lambda}\cdot x\right)\cdot\exp(-5\ln(\lambda))\cdot$
$\exp(4\ln(x) - \ln(24)) =$
$\exp\left(-\frac{1}{\lambda}\cdot x - 5\ln(\lambda) + 4\ln(x) - \ln(24)\right)$
convert all to θ: $\theta = -\frac{1}{\lambda} \to \lambda = \frac{1}{-\theta}$,
$T(x) = x,\ \ h(x) = 4\ln(x) - \ln(24),\ \ \varphi = 1,$
$A(\theta) = -5\ln(\lambda) = -5\ln\left(\frac{1}{-\theta}\right) = 5\ln(-\theta)$
since $\ln(x \le 0)$ is undefined, we must make sure $-\theta > 0$, since $\theta = -\frac{1}{\lambda}$, and $\lambda > 0,\ \theta < 0 \to -\theta > 0$.
$E[X] = A'(\theta)$ is only true when θ is ↑ in the same direction as $T(X)$, and $E[X]$ must ≥ 0, $A'(\theta)$ cannot < 0
$A'(\theta) = \frac{d}{d\theta}5\ln(-\theta) = \frac{5}{\theta}, \theta < 0, \left(\frac{5}{\theta}\right) < 0$ so this means
$E[X] = -A'(\theta) = -\frac{5}{\theta} = -\frac{5}{-\frac{1}{\lambda}} = 5\lambda$,
$Var(X) = A''(\theta)\varphi = \frac{-5}{\theta^2}\cdot 1$, Var must be ≥ 0 so $Var(X) = \frac{5}{\theta^2}\cdot 1 = \frac{5}{\left(-\frac{1}{\lambda}\right)^2} = 5\lambda^2$
$\hat{\lambda}^{MLE} = \frac{d}{d\lambda}\ln\left(\frac{x^4}{24\lambda^5}\exp\left(-\frac{x}{\lambda}\right)\right) = 0$
for MLE of a single x:
$\frac{d}{d\lambda}\ln\left(\frac{x^4}{24\lambda^5}\exp\left(-\frac{x}{\lambda}\right)\right) = 0,\ \hat{\lambda} = \frac{x}{5}$, for the MLE of all the x,
$\hat{\lambda}^{MLE} = \frac{1}{n}\cdot\frac{\sum X_n}{5} = \frac{\overline{X_n}}{5}$
We already derived that $E[X] = 5\lambda$, since $E[X] = \mu = \overline{X_n} = 1^{st}$ moment,
$\hat{\lambda}^{MM} = E\frac{X}{5} = \frac{\overline{X_n}}{5}$. $\left|\overline{X_n} - 5\right| > C_{0.05,n}$:
$\overline{X_n}$ by CLT $\to N\left(5, Var\left(\overline{X_n}\right)\right)$,
$Var\left(\overline{X_n}\right) = \frac{Var(X_i)}{n} = \frac{5\lambda_0^2}{n} = \frac{5\cdot 1^2}{n} = \frac{5}{n}$,
$C_{a,n} = q_{\frac{a}{2}}\cdot\sqrt{Var\left(\overline{X_n}\right)} = 1.96\cdot\sqrt{\frac{5}{n}}$

95^{th} quantile means the area that makes up 95% of the area under curve, can also be expressed as: $q(1-\alpha)$
critical value: T_n that marks a certain area% under curve, q_a or $q_{\frac{a}{2}}$
we use $1 - CDF$, or $1 - \varphi(T_n)$ for the right tail probabilities.
p-value: how likely the data "belongs" to H_0, the lower the p-value, the more unlikely H_0 is true. Set a α level, if $p \le a$, then H_0 might not be true, reject H_0
Significance: α =type I error, error rate for rejecting H_0 when H_0 is actually true. higher α means more likely to reject H_0.
β: type II error. Fail to reject H_0 when H_1 is actually true. Higher $\beta \to$ more likely to fail to reject H_0 when H_1 is true. $1-\beta$: power of test: prob of detecting true effect.
Workflow
1) State H_0, H_1, α
2) Compute an $\hat{\theta}$ for $H_0 : \theta_0$
3) Compute an T_n, state its distribution
4) Compute T_n's p-value
5) Find the critical value q_a (1-tail) or $q_{\frac{a}{2}}$ (2-tails) of the distribution via wolfram
6) Reject H_0 if: $T_n > q_{1-a}$, $|T_n| > q_{\frac{a}{2}}$ or for z-test: $p \le \alpha$ (p is already std z-score)

Bayesian Inference

Bayes Theorem: $\text{posterior} = \frac{\text{prior}\cdot\text{likelihood}}{\text{evidence}}$
Prior: previous belief about θ
Likelihood: Prob of observing data assuming θ is true
Evidence: Marginal that ensure the posterior integrate to 1
Posterior: Updated belief about θ given new evidence
$\pi(\theta \mid X) = \frac{\pi(\theta)\cdot L_n(X \mid \theta)}{p(X)} \propto \pi(\theta)L_n(X \mid \theta)$, remember to multiply the normalizing factor.
$p(X) = \int_\theta \pi(\theta)L_n(X \mid \theta\}$
Conjugate Prior: priors that, combined with L_n(X|θ), results in a posterior that ~ the same family of distributions as the prior. Prior: $\theta \sim B\eta(\alpha,\beta)$, Likelihood: $P(X \mid \theta) = \theta^k(1-\theta)^{n-k}$, Posterior: $\theta \mid X \sim B\eta(\alpha + k, \beta + n - k)$. $\Gamma\cdot Exp \sim \Gamma$
For proper priors: $\sum \pi(\theta) = 1$, or $\int \pi(\theta)d\theta = 1$
Improper Prior: Non-informative priors that don't integrate to 1, and are not valid distributions such as **Jeffreys** Prior: $\pi_J(\theta) \propto \sqrt{\det I(\theta)}$, invariant under reparameterization: if η=Φ(θ), we need: $\pi_J(\eta) \propto \pi_J(\theta)\cdot\left|\frac{d\theta}{d\eta}\right|$, substituting $\theta = \Phi^{-1}(\eta)$, we have $\pi_J(\eta) \propto \sqrt{\det I\left(\Phi^{-1}(\eta)\right)}\cdot\left(\Phi'\left(\Phi^{-1}(\eta)\right)\right)^{-1}$
Confidence region: $P(a \le \theta \le b \mid X) = 1 - \alpha$,
Credible Interval: $[a, b]$
Under Bayesian frameworks, the convention is to use π to denote pmf or pdf

Bayes Estimator

$\hat{\theta}$ via its $\pi(\theta \mid X)$ using a chosen loss function $L\left(\theta,\hat{\theta}\right)$: $\hat{\theta}^\pi = arg\min E\left[L\left(\theta,\hat{\theta}\right)\right] \mid X$

model, and find $L(\theta_0)$
H_1: useestimators such as MLE for all the paras of the model, and find $L\left(\hat{\theta}\right)$
LRT is defined as $\Lambda = \frac{L(\theta_0)}{L\left(\hat{\theta}\right)}$,
$T_n = -2\ln\Lambda = 2\left(\ln L\left(\hat{\theta}\right) - \ln L(\theta_0)\right) \sim \chi^2_d$, degrees of freedom = numbers of $\theta_0 - \hat{\theta}$, so if H_0 had 1 restricted λ_0, for H_1's 1 unrestricted $\widehat{\lambda}$, then $df = 1 - 0 = 1$.
$\psi_\alpha = \mathbf{1}\{T_n > q_\alpha\}$

t Test

Test if μ_1 is significantly different from μ_2, or between μ and μ_0
Assumes: iid, **X~N()**, $\sigma_1^2 \approx \sigma_2^2$, $n < 30$, uses t-table, is non-asym
t-Distribution: composed of $N(0,1)$ rvs for small n, $n < 30$.
Concept: t-test is the Z-test but instead of $Var\left(\hat{\theta}\right)$ that is asym correct, use the unbiased Var(Sample), and introduce the extra variabilities.
1-sample 1-sided: $H_0 : \mu \le \mu_0$ vs $H_1 : \mu > \mu_0$:
$T_n = \frac{\sqrt{n}\left(\overline{X_n} - \mu_0\right)}{\sqrt{\widetilde{S_n}}} \sim t_{n-1}$, $\psi_\alpha = \mathbf{1}\{T_n > q_\alpha\}$
1-sample 2-sided: $H_0 : \mu = \mu_0$ vs $H_1 : \mu \ne \mu_0$
$T_n = \frac{\sqrt{n}\overline{X_n}}{\sqrt{\widetilde{S_n}}} = \frac{\sqrt{n}\frac{\overline{X_n} - \mu_0}{\sigma}}{\sqrt{\frac{\widetilde{S_n}}{\sigma^2}}} \sim t_{n-1}$,
$\psi_\alpha = \mathbf{1}\left\{|T_n| > q_{\frac{\alpha}{2}}\right\}$
2-samples (Welch's Test)
special case of 2-sampled t-test where, $\sigma_1^2 \ne \sigma_2^2$ (heteroscedasticity), and $n_1 \ne n_2$
$\overline{X_n} \sim N\left(\mu_X, \frac{\sigma_X^2}{n}\right)$ and $\overline{Y_m} \sim N\left(\mu_\Upsilon, \frac{\sigma_\Upsilon^2}{m}\right)$
$\frac{\overline{X_n} - \overline{Y_m} - (\mu_X - \mu_\Upsilon)}{\sqrt{\frac{\tilde{\sigma}_X^2}{n} + \frac{\tilde{\sigma}_Y^2}{m}}} \sim t_N$,
$N = \frac{\left(\frac{\tilde{\sigma}_X^2}{n} + \frac{\tilde{\sigma}_Y^2}{m}\right)^2}{\frac{\tilde{\sigma}_X^4}{n^2(n-1)} + \frac{\tilde{\sigma}_Y^4}{m^2(m-1)}} \ge \min(n, m)$

Bayes Estimator Example

1) $\pi(p = 0.4) = 0.2, \pi(p = 0.7) = 0.8$
2) Compute all likelihoods $L(X \mid p)$:
$L(X = 3 \mid p = 0.4) = \binom{6}{3}\cdot 0.4^3\cdot(1 - 0.4)^{6-3}$
$= 0.27648, L(X = 3 \mid p = 0.7) = 0.18522$
3) Compute all posteriors $\pi(p \mid X)$:
$\pi(p = 0.4 \mid X = 3) = \pi(p = 0.4)\cdot L(X = 3 \mid p = 0.4)$
$= 0.2\cdot 0.27648 = 0.055296, \pi(p = 0.7 \mid X = 3) = 0.148176$
4) Normalize $\pi(p \mid X)$: $\frac{\pi(p = 0.4 \mid X = 3)}{\sum \pi(p \mid X)}$
$= 0.27176, \pi(p = 0.7 \mid X = 3) = 0.72823$
normalized $\pi(p \mid X)$ should sum to 1.
5) $\widehat{p}^{Bayes} = E[p \mid X = 3] = \sum p_i\cdot\pi(p \mid X)$
$= (0.4\cdot 0.27176) + (0.7\cdot 0.72823) = 0.618$
Even though $p \in \{0.4, 0.7\}$, $\widehat{p}^{Bayes}$ thinks that $E[p \mid X] = 0.618$ means the outcome is closer

e limits
$\lim_{n\to\infty}\left(1 - \frac{t}{n}\right)^n = e^{-t}$, $\lim_{n\to\infty}\left(1 + \frac{t}{n}\right)^n = e^t$
Indicators
$I_A = 1$ if $P(A)$, $I_A = 0$ if $P(A^c)$, $E[I_A] = P(A)$

Wolfram Syntax

Distros as objects: NormalDistribution[μ,σ]
⚠beware σ is std dev, not var!
Commands: PDF[NormalDistribution[μ,σ],x], plots can be added on top
Plot[PDF[NormalDistribution[2, 3],x],{x,-5,10}],
Mean[], Moment[], **Expected Value**: $EV[f(x)]$
Quantile[NormalDistribution[0,1], 0.975] for the Z-score of $a = 0.025$ of a standard normal, aka **inverse CDF**
Probability $x \le 1.96$, NormalDistribution[0,1] for the area under curve.
Quantile[StudentTDistribution[4],1-a] for the critical value of the area to the left of α
CDF[NormalDistribution[0, 1], 1.96] for area under curve upto x=1.96 of a standard normal
MLE: all components in a single step: $\frac{d}{d\theta}\ln\left(\prod f_\theta(x)\right) = 0$, look for "real solutions" of θ or x. Note: use x for $\sum X_i$. ⚠ Watch out for similar chars.

Delta Method

Use CLT to $\approx Var\left(\hat{\theta}\right)$ of re-parameterized $\hat{\theta}$. Asym $V\left(\hat{\theta}\right) = Var\left(\hat{\theta}\right)\cdot(g'(\theta))^2$, this accounts for the change in variability induced by the transformation.
`g(θ)` must be continous differentiable
$\sqrt{n}\left(\hat{\theta} - \theta\right) \to N\left(0, \sigma^2\right)$
$\sqrt{n}\left(g\left(\hat{\theta}\right) - g(\theta)\right) \to N\left(0, (g'(\theta))^2\cdot\sigma^2\right)$
Example: We want to estimate $\sqrt{\mu}$ instead of the mean μ. We define $g(x) = \sqrt{x}$, and use $g\left(\overline{X}\right) = \sqrt{\overline{X}}$ as an estimator for $\sqrt{\mu}$. By delta method, we have: $\sqrt{n}\left(g\left(\overline{X}\right) - g(\mu)\right) \to N\left(0, (g'(\mu))^2\cdot\sigma^2\right)$.
Next, given $g(\mu) = \sqrt{\mu}$, $g'(\mu) = \frac{1}{2\sqrt{\mu}}$, then,
$Var\left(\sqrt{\overline{X}}\right) \approx \frac{(g'(\mu))^2\cdot\sigma^2}{n} = \frac{\left(\frac{1}{2\sqrt{\mu}}\right)^2\cdot\sigma^2}{n} = \frac{\sigma^2}{4n\mu}$

Delta Method Example 2

1) Given $Y_i \sim Ber\left(\theta^t\right), \overline{Y_n}$ is an estimator for θ^t, and $\hat{\theta} = \overline{Y_n}^{\frac{1}{t}}$
2) We need $Var\left(\hat{\theta}\right)$ as $n \to \infty$
3) $Var(Y_i) = \theta^t\cdot\left(1 - \theta^t\right), Var\left(\overline{Y_n}\right) = \frac{\theta^t\left(1 - \theta^t\right)}{n}$
4) Delta Method: Let $g(x) = x^{\frac{1}{t}}$
$g'(x) = \left(\frac{1}{t}\right)x^{\frac{1}{t}-1}, \left(g'\left(\theta^t\right)\right)^2 = \left(\frac{1}{t^2}\right)\theta^{2-2t}$
5) $Var\left(\hat{\theta}\right) \approx \left(g'\left(\theta^t\right)\right)^2\cdot Var\left(\overline{Y_n}\right)$:
6) $Var\left(\hat{\theta}\right) = \left(\frac{\theta^t\left(1 - \theta^t\right)}{n}\right)\cdot\left(\frac{1}{t^2}\right)\cdot\theta^{2-2t}$

Log-L: $l(\theta;x)=\sum_{i=1}\left[\ln h(X_i)+\frac{\quad}{\varphi}\right]$

Canonical Families:
$N(\mu,\sigma^2), Ber(p), Poi(\lambda), Exp(\lambda), \Gamma(\alpha,\beta), \text{Binom}(k,p)$

Linear Regression

$Y=a+bX+\varepsilon$. Y: dependent/response, X: Independent/Predictor, a: intercept, b: slope, ε: error.
Goal: Fit a line $y=a+bx$, and find a,b values that minimize some loss.
OLS (vertical): $\sum(Y-a-bX)^2$: $\hat{a}=\overline{Y}-\hat{b}\overline{X}$, $\hat{b}=\frac{Cov(X,Y)}{Var(X)}=\frac{\sigma_{XY}}{\sigma^2\_X}=\frac{\overline{XY}-\overline{X}\cdot\overline{Y}}{\frac{\sum X^2}{n}-\left(\overline{X}\right)^2}$

wolfram: OLS: $\text{fit linear}(x,y),(x,y)..$, $Cov(X,Y)$: $\text{covariance}(x,x..)(y,y..)$, this gives unbiased, must $\cdot\frac{n-1}{n}$ to get the biased σ_{XY}, same goes for $\text{variance}(x,x..)$

horizontal distance: $\sum\left(X-\frac{Y-a}{b}\right)^2$

$\frac{1}{b}=\frac{\sigma_{XY}}{\sigma_Y^2}\to b=\frac{\sigma_Y^2}{\sigma_{XY}}$

Min(X) & Max(X)

$P(\max>x)=1-P(\max<x)=1-[P(X_i<x)]^n$

$P(\max<x)=\prod_{i=1}^{n}P(X_i<x)=[P(X_i<x)]^n$

$P(\min>x)=[P(X_i>x)]^n=[1-P(X_i<x)]^n$

$P(\min>x)=\prod_{i=1}^{n}P(X_i>x)=[P(X_i>x)]^n$

$P(\min<x)=1-P(\min>x)$

Total Variation Distance

Measures max distance between two distributions $P_\theta, P_{\theta'}$

$TV\left(P_\theta,P_{\theta'}\right)=\frac{1}{2}\sum_{x\in E}\left|p_\theta(x)-p_{\theta'}(x)\right|$

$TV\left(P_\theta,P_{\theta'}\right)=\frac{1}{2}\int_{-\infty}^{\infty}\left|f_\theta(x)-f_{\theta'}(x)\right|dx$

E is the joint support of $P_\theta, P_{\theta'}$.

Properties

symmetric: $TV\left(P_\theta,P_{\theta'}\right)=TV\left(P_{\theta'},P_\theta\right)$

positive: $0\le TV\le 1$

definite: if $TV\left(P_\theta,P_{\theta'}\right)=0$ then $P_\theta=P_{\theta'}$

triangle inequality:

$TV\left(P_\theta,P_{\theta'}\right)\le TV\left(P_\theta,P_{\theta''}\right)+TV\left(P_{\theta''},P_{\theta'}\right)$

if $TV=1: then P_\theta, P_{\theta'}$ disjoint,
if $TV=0: then P_\theta, P_{\theta'}$ same.

Multiple Hypothesis Testing

Testing Multi-θ α↑, we must control family-wise err=P(at least 1 type_I)≤α

Bonferroni's test: Divide α by # of tests m: Reject H_i if $p_i\le\frac{\alpha}{m}$

Example: $m=100$ at $\alpha=0.05$:

$\text{FWER}\approx 1-P(type_I=0)=1-(1-0.05)^{100}\approx 0.994$

Very restrictive for large m →lower β
FDR=fraction of type_I in all sig≤α

Bonferroni corr: reject if $m\cdot\text{p-value}\le\alpha$, controls FWER, but very conservative.

Holm-Bonferroni corr: sort p-values, reject smallest p_k: $p_k>\frac{\alpha}{m-k+1}$

Bonferroni-Hochberg corr: sort p-values, threshold: $p_k\le\frac{k\cdot\alpha}{m}$, $k=1,..,m$

Log-Likelihood

$l(x\mid\theta)=\ln\left(\prod_{i=1}^{n}f_\theta(x)\right)=\sum_{i=1}^{n}\ln\left(f_\theta(x)\right)$

Method of Moments Estimator

Analyze which aspect the parameter θ represents, then choose a moment that also represent the aspect as the

$\hat{\theta}^{MM}=\overline{X}_n^k=\frac{\sum X_i^k}{n}$

by LLN: $\hat{\theta}_k\to E_\theta\left[X^k\right]$

by CLT: $\sqrt{n}\left(\hat{\theta}-\theta\right)\to N(0,\Gamma(\theta))$

$\Gamma(\theta)$: asym covariance matrix.

$\hat{\theta}_1=\mu=E[X]=\frac{\sum X_i}{n}=\overline{X}_n$,

$\hat{\theta}_2=\mu^2+\sigma^2=E\left[X^2\right]=\frac{\sum X_i^2}{n}$,

$\sigma^2=E\left[(X-\mu)^2\right]=\frac{\sum\left(x_i-\overline{X}\right)^2}{n}$ biased,

$\mu'_k=E\left[X^k\right]$, $\mu_k=E\left[(X-\mu)^k\right]$

KL Divergence

Measures the difference between 2 distributions:

$KL\left(P_\theta,P_{\theta'}\right)=\sum p_\theta(x)\log\left(\frac{p_\theta(x)}{p_{\theta'}(x)}\right)$

$KL\left(P_\theta,P_{\theta'}\right)=\int f_\theta(x)\log\left(\frac{f_\theta(x)}{f_{\theta'}(x)}\right)dx$

Properties

not symmetric:

$KL\left(P_\theta,P_{\theta'}\right)\ne KL\left(P_{\theta'},P_\theta\right)$

not negative: $KL\left(P_\theta,P_{\theta'}\right)\ge 0$

definite: $KL\left(P_\theta,P_{\theta'}\right)=0i$ if $P_\theta=P_{\theta'}$

KS Test:

Goodness of fit by comparing empirical CDF w/ CDF. Do not require bins. Provide numerical $T_n=\sqrt{n}\sup\left|F_n(t)=F_0(t)\right|$, p-value: $P(Z>T_n)$ assume H_0 is true, reject H_0: $T_n>q_\alpha$

KL Test:

Similar to KS test but θ are estimated. More likely to reject than KS. If θ is given, not suitable. Test if Gaussian: $\sup\left|F_n(t)-\varphi_{\hat{\mu},\hat{\sigma}^2}(t)\right|$

Geometric Distribution
$Geom(p)=$ similar to exp() but discrete, # of trials until 1^{st}
$P(X=k)=(1-p)^{k-1}p$, $k=1,2\ldots$.

Significance Tests

is $j^{th}X_j$ significant to Y?
$H_0:\beta_j=0$ vs $H_1:\beta_j\ne 0$

$\gamma_j: j^{th}$ diag coeff $\left(\mathbb{X}^T\mathbb{X}\right)^{-1}$ $(\gamma_j>0)$

$T_n^{(j)}=\frac{\hat{\beta}_j-\beta_j}{\sqrt{\hat{\sigma}^2\gamma_j}}\sim t_{n-p}$, $\hat{\sigma}^2=\frac{\left\|Y-\mathbb{X}\hat{\beta}\right\|^2}{n-p}$

Rejection Region:

$R_{j,\alpha}=\left\{\left|T_n^{(j)}\right|>q_{\frac{\alpha}{2}}\left(t_{n-p}\right)\right\}$

function $L(\theta,\hat{\theta})$: $\hat{\theta}=arg\min_{\hat{\theta}} E\left[L(\theta,\hat{\theta})\right]\mid X$

Bayes is the posterior **mean:**

$\hat{\theta}^\pi=E[\theta\mid X]=\int\theta\cdot\pi(\theta\mid X)d\theta$, or $\sum\theta\cdot\pi(\theta\mid X)$, if using $\propto$, divide each $\pi(\theta\mid X)$ by $\sum\pi(\theta\mid X)$ aka LMS, or conditional expectation.

MAP is posterior **mode:** $\hat{\theta}^{MAP}=arg\max_\theta\pi(\theta\mid X)$

Must use normalized $\pi(\theta\mid X)$, not the $\propto\pi(\theta)L(X\mid\theta)$ version without $p(X)$.
Or use the mean formula of known distribution types. asym $Var\left(\hat{\theta}\right)=\frac{1}{I(\theta)}$.

LMS Error: Error=Bayes-Asym_Value,
$E[Err\text{ or }\mid X]=0$, $Cov\left(Err\text{ or },\hat{\theta}\right)=0$,
$Var(\theta)=Var\left(\hat{\theta}\right)+Var(Err\text{ or })$.

Conditional MSE of Bayes is the posterior Var under quad loss: $E\left[\left(\theta-\hat{\theta}\right)^2\mid X\right]=Var(\theta\mid X)$

Multivariate Linear Regression

$\vec{Y}=\mathbb{X}\vec{\beta}^\star+\vec{\varepsilon}$, $\vec{\beta}\in\mathbb{R}^p$, $\vec{Y}\in\mathbb{R}^n$, $\mathbb{X}\in\mathbb{R}^{n\times p}$.

$X^T\beta=\mu(x)=E[Y\mid X=x]=\int y\cdot h(y\mid x)dy$

LSE $\hat{\vec{\beta}}^{Bayes}$: $\hat{\vec{\beta}}=arg\min\left\|\vec{Y}-\mathbb{X}\vec{\beta}\right\|^2$, solution:

$\hat{\vec{\beta}}=\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\mathbb{X}^TY$, must be full rank:
$\text{Rank}(X)=p$, and $n\ge p$.

Assumptions: Design Matrix $\mathbb{X}$ deterministic, full rank, $\vec{\varepsilon}\sim N\left(0,\sigma^2I_n\right)$

iid: $\Rightarrow Y\sim N_n\left(\mathbb{X}\beta^\star,\sigma^2I_n\right)$, $\Rightarrow I(\beta)=\frac{1}{\sigma^2}\mathbb{X}^T\mathbb{X}$

Properties of LSE
LSE=MLE in homoscedastic Gaussian

Distro of $\hat{\beta}\sim N_p\left(\beta^\star,\sigma^2\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\right)$, is asym.

Quad Risk: $E\left[\left\|\hat{\beta}-\beta\right\|^2\right]=\sigma^2\text{trace}\left(\left(\mathbb{X}^T\mathbb{X}\right)^{-1}\right)$

Prediction Error: $E\left[\left\|Y-\mathbb{X}\hat{\beta}\right\|^2\right]=\sigma^2(n-p)$

Unbiased $\hat{\sigma}^2=\frac{\left\|Y-\mathbb{X}\hat{\beta}\right\|^2}{n-p}=\frac{1}{n-p}\sum\hat{\varepsilon}^2$

$(n-p)\frac{\hat{\sigma}^2}{\sigma^2}\sim\chi^2_{n-p}$, $\hat{B},\hat{\sigma}^2$ orthogonal and indie

Generalized Linear Models "GLM"

Link μ of response Y linearly to predictor X: $\mu_i=b'(\theta_i)$. $\mu=E[Y\mid X]$, via Link fn:
$g(\mu_i)=\theta=a+bX=X_i^T\beta$, inverse: $\mu=g^{-1}(\theta)$.
$\hat{\beta}$: asym normal, find $\boldsymbol{\beta}$ via MLE.

$N:g(\mu)=\mu=X^T\beta$, $\text{Binom}:g(\mu)=\ln\left(\frac{\mu}{n-\mu}\right)$

$Exp:g(\mu)=-\mu^{-1},\mu=-\left(X^T\beta\right)^{-1}$,

$\text{Gamma}:g(\mu)=\mu^{-2},\mu=\left(X^T\beta\right)^{-\frac{1}{2}}$,

$Poi:g(\mu)=\ln(\mu),\mu=\exp\left(X^T\beta\right)$,

$Ber:g(\mu)=\ln\left(\frac{\mu}{1-\mu}\right),\mu=\frac{1}{1+\exp\left(-X^T\beta\right)}$,

Bonferroni's Test

test group β is sig at FWER≤ α, non-asym.
$H_0:\beta_j=0\forall j\in S$ where $S\subseteq\{1,\ldots,p\}$
$H_1:\exists j\in S$ where $\beta_j\ne 0$

$\psi=\mathbf{1}\left\{\frac{\max\left(\left|\hat{\beta}_1\right|,\left|\hat{\beta}_2\right|,\ldots\right)}{\sqrt{Var\left(\hat{\beta}_j\right)}}>q_{\frac{\alpha}{2k}}\right\}$

$E[p\mid X]=0.615$ means the outcome is closer to 0.7 than 0.4, $\hat{p}^{MAP}=0.7$ because MAP is the mode and $P(p=0.7)>P(p=0.4)$.

Categorical Likelihood

Are Zodiac signs~ $\text{Unif}(a,b)$? $p_0=\left(\frac{1}{12},\frac{1}{12}\ldots\right)$, prob of a $X_i\in j$: $p_j=j(P(X=a_j))$, # of $X_i\in j$: $N_j=n\{X_i=a_j\}$, likelihood:

$L_n=p_1^{N_1}\cdot p_2^{N_2}\ldots=\prod_{j=1}^{k}p_j^{N_j}$ with $\vec{p}=(p_1,p_2,\ldots,p_k)$,

$\vec{p}^{MLE}$: $\hat{p}_j=\frac{N_j}{n}$ where $\sum p_j=1$, for a single $X_i J$: $P(X=a_j)=\prod_{j=1}^{k}p_i^{1(X_i=a_j)}$

Chi Squared Distribution

χ_k^2: $\sum$ of k squared $N(0,1)$ variables, with degrees of freedom= k

$E[X]=k$, $Var(X)=2k$, $f(x)=\frac{1}{2^{\frac{k}{2}}\Gamma\left(\frac{k}{2}\right)}x^{\frac{k}{2-1}}e^{-\frac{x}{2}}$

$\chi^2\sim\Gamma\left(\frac{k}{2},2\right)$, $\chi^2\sim N\left(k,\sqrt{2k}\right)$ for large k.

As n grows, χ_1^2 (composed from a single normal) 95^{th} quantile value decreases, and the distribution is scaled by
$\text{asym}\cdot q_a+\frac{\text{any }q_a-\text{asym }q_a}{\sqrt{n}}$

Chi Squared Test

χ^2 test is a goodness-of-fit test (match distribution, variable independence) that groups categorical data into bins, then compare observed data freq w/ expected data freq, requires counts in each bin > 5.

$H_0:\vec{p}=\vec{p}^0$ vs. $H1:\vec{p}\ne\vec{p}^0$

$T_n=n\sum^k\left[\frac{\left(\hat{p}_j-p_j^0\right)^2}{p_j^0}\right]\to\chi^2_{k-1}$, count of j in n: $p_j=\frac{n_j}{n}$, k: # of categories.

$H_0:p\in\{\text{Bin}(k,\theta)\}$ vs $H_1:p\notin\{\text{Bin}(k,\theta)\}$

$T_n=n\sum_{j=0}^{k}\frac{\left(\frac{N_j}{n}-f_{\hat{\theta}}(j)\right)^2}{f_{\hat{\theta}}(j)}\to\chi^2_{(k+1)-d-1}$, $\hat{\theta}$: MLE,

$f_{\hat{\theta}}(j)=\binom{k}{j}\hat{\theta}^j\left(1-\hat{\theta}\right)^{k-j}$, d: # of para,
$k-d-1$ when start at j=1

Optimization

global extremes could be on the supports (end points), if distribution defined on closed range.
solve $h'(x)=0$ for critical points.
min/max critical points
$h''(x)\le 0$: concave, max, $h'(x)\downarrow$
$h''(x)<0$: str concave, global max, $h'(x)\downarrow$
$h''(x)\ge 0$: convex, min, $h'(x)\uparrow$
$h''(x)\ge 0$: str convex, global min, $h'(x)\uparrow$
Multivariate min/max
Gradient is the vector of partial deri:
$\nabla h(\theta)=\left(\frac{dh}{d\theta_1},\ldots\right)$, Hessian $Hh(\theta)$ is the matrix of 2^{nd} partial deri:
$X^THh(\theta)X\le 0$: concave, max, vice versa
If diagonal of Hessian are positive and Hessian symmetric: convex, minimum

$\left(\quad\right)\left(\quad n\quad\right)\left(t^2\right)$

$=\frac{\theta^{2-t}\left(1-\theta^t\right)}{t^2}$ We can remove n to show that asym variance has large n already.

Normal-like Functions

$f_X(x)=c\exp\left(-\left(\alpha x^2+\beta x+\gamma\right)\right)$, complete the square: $-\alpha\left(x+\frac{\beta}{2\alpha}\right)^2+c$.

therefore: $\mu=-\frac{\beta}{2\alpha}$, $\sigma^2=\frac{1}{2\alpha}$.

peak of $f_X(x)$ = min of $-\left(\alpha x^2+\beta x+\gamma\right)$:

$\frac{d}{dx}\left(-\alpha x^2-\beta x-\gamma\right)=-2\alpha x-\beta$, $-2\alpha x-\beta=0$,

$x=\mu=-\frac{\beta}{2\alpha}$.

When posterior unimodal & symmetric, such as~ $N()$: $\hat{\theta}^{MAP}=\hat{\theta}^{LMS}=\mu$

Covariance Matrix

$$\Sigma=\begin{pmatrix}Cov(X,X) & Cov(X,Y)\\ Cov(Y,X) & Cov(Y,Y)\end{pmatrix}$$

$=E[(X-E[X])(Y-E[Y])]$
For single vector X: $Var(\mathbf{X})=Cov(\mathbf{X})$
Linear:
$Cov(\mathbf{AX}+\mathbf{B})=Cov(\mathbf{AX})=\mathbf{A}Cov(\mathbf{X})\mathbf{A}^T=\mathbf{A}\Sigma\mathbf{A}^T$

Multivariate Gaussian vector: defined by mean vector $\boldsymbol{\mu}$ and Σ

$f_X(x)=\frac{1}{\sqrt{(2\pi)^d\det\Sigma}}\exp\left(-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right)$

Multivariate CLT
$X_i\sim R^d$ $E[\mathbf{X}_i]=\boldsymbol{\mu}$ $Cov(\mathbf{X}_i)=\Sigma$

Multivariate Delta Method
$\sqrt{n}(g(T_n)-g(\theta))\to N\left(0,\nabla g(\theta)^T\Sigma\nabla g(\theta)\right)$

Miscellaneous Theorems

Markov Inequality: $X\ge 0$ and $a\ge 0$, provides upper bound on $P(X\ge a)$.

$P(X\ge a)\le\frac{E[X]}{a}$

Chebyshev Inequality: upper bound on $P(X$ deviates from $\boldsymbol{\mu}$ by > **c** std dev)

$P(|X-\mu|>c)\le\frac{\sigma^2}{c^2}$

Slutsky's Theorem: Relationships between a rv that → rv and a rv that → constant.
Given $T_n\to T$ and $U_n\to c$, then,
$T_n+U_n\to T+c$, $T_n\cdot U_n\to T\cdot c$, $\frac{T_n}{U_n}\to\frac{T}{c}$

Convergence: $\to_d$: In distribution, T_n's distro → T's distro.
$\to_p$: In probability, $U_n\to c$: for any $\text{error}>0$, $P(|U_n-c|>\varepsilon)\to 0$

Cochran's Theorem: Relationship between S_n,χ^2, typically in ANOVA
$\frac{nS_n}{\sigma^2}\sim\chi^2_{n-1}$ or $S_n\sim\frac{\sigma^2}{n}\chi^2_{n-1}$

Donsker's Theorem: connects empirical CDF to Brownian bridge (A Gaussian Process). if F is cont CDF:
$\sqrt{n}\sup|F_n(t)-F(t)|\to\sup_{t\in[0,1]}|B(t)|$

Moivre Laplace Correction: using discrete rv to ≈ conti rv via CLT.

$S_n\sim\text{Bin}(n,p):P(S_n\le k)\approx P\left(Z\le\frac{k-np+0.5}{\sqrt{np(1-p)}}\right)$