Santucci's Stats 200 Notes

Basic Probability

Cauchy-Schwarz	Markov	Chebyschev
$\lvert E(XY)\rvert \le \sqrt{E(X^2)E(Y^2)}$	$P(X \ge t) \le \frac{E[X]}{t}$	$P(\lvert X-\mu\rvert \ge t) \le \frac{\sigma^2}{t^2}$

Conditional Expectation $E[g(X)|Y=y] = \int_{-\infty}^{\infty} g(x) f^{(X|Y)}(x|y)dx$

Conditional Variance $Var[g(X)|Y=y] = E[\{g(X)\}^2|Y=y] - \{E[g(X)|Y=y]\}^2$

Total Expectation $E[g(X)] = E\{E[g(X)|Y]\}$

Total Variance $Var(X) = E[Var(g(X)|Y)] + Var(E[g(X)|Y])$

Convergence Concepts

Convergence in Probability $\{X_n : n \ge 1\}$ converges in probability to X if $\forall \epsilon > 0$: $Pr(|X_n - X| > \epsilon) \to 0$.

Convergence in Distribution $\{X_n : n \ge 1\}$ converges in distribution to X if $F^{(X_n)}(x) \to F^{(X)}(x)$ *at every point where F is continuous.*

Thrm. If $X_n \xrightarrow{P} X$, then $X_n \xrightarrow{D} X$.

Thrm. Let $\alpha \in \mathbb{R}$ be a constant. Then $X_n \xrightarrow{P} \alpha \iff X_n \xrightarrow{D} \alpha$.

Showing Convergence in Probability Options: show (1) directly through definition, (2) if convergence to a constant, try showing convergence in distribution, or (3) use thrm.: if $E[X_n] \to \alpha \in \mathbb{R}$ and $Var(X_n) \to 0$, $\implies X_n \xrightarrow{P} \alpha$.

Showing Convergence in Distribution Options: show (1) Convergence in Probability, (2) Convergence in Distribution through CDF's, or (3) CLT [requires i.i.d. and sums/average].

Continuous Mapping Theorems

Thrm. If $X_n \xrightarrow{P} \alpha$ for some constant $\alpha \in \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous at α, then $g(X_n) \xrightarrow{P} g(\alpha)$
(This is also true if $X_n \xrightarrow{D} \alpha$, using the above thrm.)

Thrm. If $X_n \xrightarrow{P} X$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \xrightarrow{P} g(X)$

Thrm. If $X_n \xrightarrow{D} X$ and $g : \mathbb{R} \to \mathbb{R}$ is continuous, then $g(X_n) \xrightarrow{D} g(X)$

Slutsky's Theorem

If $X_n \xrightarrow{D} X$ and $Y_n \xrightarrow{P} \alpha$, where $\alpha \in \mathbb{R}$ is a constant, then $X_n + Y_n \xrightarrow{D} X + \alpha$ and $X_n Y_n \xrightarrow{D} aX$.

Weak Law of Large Numbers

Let $\{X_n : n \ge 1\}$ be a sequence of i.i.d. R.V.'s with $E[|X_1|] < \infty$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then, $\bar{X}_n \xrightarrow{P} E[X_1]$.

Proof Let $X_1, ...X_n$ be i.i.d (μ, σ^2), with $\sigma^2 < \infty$, Chebyschev implies $Pr(|\bar{X}_n - \mu| < \epsilon) \le 1 - \frac{\sigma^2}{n\epsilon^2}$. Hence, $\lim n \to \infty\, Pr(|\bar{X}_n - \mu| < \epsilon) = 1$. □

Central Limit Theorem

The asymptotic distribution of an average of i.i.d. R.V.'s is a normal distribution, regardless of the individual random variables themselves.

Thrm. Let $\{X_n : n \ge 1\}$ be a sequence of i.i.d. R.V.'s with $Var(X_1) < \infty$. Let $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. Then

$$\frac{\sqrt{n}(\bar{X}_n - \mu)}{\sigma} \xrightarrow{D} \mathcal{N}(0,1)$$

where $\mu = E[X_1]$ and $\sigma^2 = Var(X_1)$.

Delta Method

Basic idea: $g(Y_n) - g(\alpha) \approx g'(\alpha)(Y_n - \alpha)$

Thrm. Let $\{Y_n : n \ge 1\}$ be a sequence of random variables such that $\sqrt{n}(Y_n - \alpha) \xrightarrow{D} Z$ for some random variable Z and some constant $\alpha \in \mathbb{R}$. Let $g : \mathbb{R} \to \mathbb{R}$ be *continuously differentiable* at α. Then,

$$\sqrt{n}[g(Y_n) - g(\alpha)] \xrightarrow{D} g'(\alpha)Z$$

Proof Formally, $\sqrt{n}[g(Y_n) - g(\alpha)] = g'(Y_n*)\sqrt{n}(Y_n - \alpha)$ for some Y_n* between α and Y_n. Note that for any $\epsilon > 0$, $Pr(|Y_n * - \alpha| > \epsilon) \le Pr(|Y_n - \alpha| > \epsilon) \xrightarrow{P} 0$ since $Y_n \xrightarrow{P} \alpha$ (through WLLN). Then, $Y_n* \xrightarrow{P} \alpha$, so $g'(Y_n*) \xrightarrow{P} g'(\alpha)$ by our first continuous mapping theorem. Since $\sqrt{n}[Y_n - \alpha] \xrightarrow{D} Z$, the result follows by Slutsky's theorem. □

Random Vectors

Expectation $E[\boldsymbol{X}] = [E[X_1], ..., E[X_n]]$

Variance $Var(\boldsymbol{X}) = E[\{\boldsymbol{X} - E[\boldsymbol{X}]\}\{\boldsymbol{X} - E[\boldsymbol{X}]\}^\intercal] = E[\boldsymbol{X}\boldsymbol{X}^\intercal] - E[\boldsymbol{X}]E[\boldsymbol{X}]^\intercal$

Linearity $E[\boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{X} + \boldsymbol{C}\boldsymbol{Y}] = \boldsymbol{\alpha} + \boldsymbol{B}E[\boldsymbol{X}] + \boldsymbol{C}E[\boldsymbol{Y}]$
$Var(\boldsymbol{\alpha} + \boldsymbol{B}\boldsymbol{X}) = \boldsymbol{B}Var(\boldsymbol{X})\boldsymbol{B}^\intercal$

Multivariate Normal Distribution

Definition Let $\boldsymbol{Z}$ be a random vector with $\boldsymbol{\theta} = E[\boldsymbol{Z}]$ and $\boldsymbol{V} = Var(\boldsymbol{Z})$. $\boldsymbol{Z}$ is called *multivariate normal*, denoted $\boldsymbol{Z} \sim N_p(\boldsymbol{\theta}_p, \boldsymbol{V}_p) \iff \boldsymbol{\alpha}^\intercal\boldsymbol{Z}$ has a univariate normal distribution for all $\alpha \in \mathbb{R}^p$. The following properties hold:

PDF If $\boldsymbol{V}$ is non-singular (invertible), then

$$f(\boldsymbol{z}) = \frac{1}{(2\pi)^{p/2} det \boldsymbol{V}^{1/2}} \exp\left[-1/2(\boldsymbol{z} - \boldsymbol{\theta})^\intercal \boldsymbol{V}^{-1}(\boldsymbol{z} - \boldsymbol{\theta})\right]$$

where det denotes determinant.

Independence $Z_i \perp\!\!\!\perp Z_j \iff V_{ij} = Cov(Z_i, Z_j) = 0$.

Standard Normal Let $\boldsymbol{0}_p$ Denote a zeros vector length p, and $\boldsymbol{I}_p$ denotes the $p \times p$ identity matrix. $N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$ is called the *p-variate standard normal distribution.*

Lemma Let $\boldsymbol{A}$ be a $p \times p$ matrix that is orthogonal ($\boldsymbol{A}\boldsymbol{A}^\intercal = \boldsymbol{A}^\intercal\boldsymbol{A} = \boldsymbol{I}_p$), and let $Z \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$. Then $\boldsymbol{A}\boldsymbol{Z} \sim \mathcal{N}_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$.

Proof For any vector $\boldsymbol{b} \in \mathbb{R}^p$, the random vector $\boldsymbol{b}^\intercal\boldsymbol{A}\boldsymbol{Z} = (\boldsymbol{A}^\intercal\boldsymbol{b})^\intercal\boldsymbol{Z}$ has a univariate normal since $\boldsymbol{Z}$ is multivariate normal. Then $\boldsymbol{A}\boldsymbol{Z}$ is multivariate normal. Now simply note that $E[\boldsymbol{A}\boldsymbol{Z}] = \boldsymbol{A}E[\boldsymbol{Z}] = \boldsymbol{0}_p$ and that $Var(\boldsymbol{A}\boldsymbol{Z}) = \boldsymbol{A}\boldsymbol{I}_p\boldsymbol{A}^\intercal = \boldsymbol{A}\boldsymbol{A}^\intercal = \boldsymbol{I}_p$ □

Sample Variance

Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $n \ge 2$. $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$ where $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^n}{n})$ and

$$S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2 = \frac{1}{n-1}\left[\sum_{i=1}^n X_i^2 - (\bar{X}_n)^2\right]$$

Chi-Squared Distribution

Let $\boldsymbol{Z} \sim N_p(\boldsymbol{0}_p, \boldsymbol{I}_p)$, then $\boldsymbol{Z}^\intercal\boldsymbol{Z} = \sum_{i=1}^p Z_i^2$ is called a *chi-squared* distribution with p-degrees of freedom, with expectation p and variance $2p$. Recall: $Var(Z_i) = 1 = E[Z_i^2]$

Lemma The χ_1^2 distribution is the Gamma(1/2, 1/2) distribution.

Lemma Let $U_1, ..., U_n$ be independent with $U_i \sim Gamma(\alpha_i, \beta)$ for each $i \in \{1, ...n\}$. Then $\sum_{i=1}^n U_i \sim Gamma(\sum_{i=1}^n \alpha_i, \beta)$.

Lemma $\chi_p^2 \sim Gamma(p/2, 1/2)$

Joint Dist.: Sample Mean/Variance

Thrm. Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $n \ge 2$. Then $\bar{X}_n \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$ and $\frac{(n-1)S^2}{\sigma^2} \sim \chi_{n-1}^2$. Further, $\bar{X}_n$ and S^2 are independent.

Proof Sufficient to prove for $\mu = 0$ and $\sigma^2 = 1$. Let $\boldsymbol{X} = (X_1, ..., X_n) \sim N_n(\boldsymbol{0}_n, \boldsymbol{I}_n)$. Now let A be and orthogonal $p \times p$ matrix, for which all the elements in the first row are $\frac{1}{\sqrt{n}}$, constructed via Graham-Schmidt. Let $\boldsymbol{Y} = (Y_1, ..., Y_n) = \boldsymbol{A}\boldsymbol{X}$. By a previous lemma, $Y \sim N_n(\boldsymbol{0}_n, \boldsymbol{I}_n)$, so the sum of squares of its last $n-1$ elements is $\sum_{i=2}^n Y_i^2 \sim \chi_{n-1}^2$. Note that the first element is $Y_1 = \sqrt{n}\bar{X}_n$, so we may write: $\sum_{i=2}^n Y_i^2 = \sum_{i=1}^n Y_i^2 - Y_1^2 = \boldsymbol{X}^\intercal\boldsymbol{A}^\intercal\boldsymbol{A}\boldsymbol{X} - n(\bar{X}_n)^2 = \boldsymbol{X}^\intercal\boldsymbol{X} - n(\bar{X}_n)^2 = \sum_{i=1}^n X_i^2 - n(\bar{X}_n)^2 = (n-1)S^2$ Finally, note that $Y_1, ..., Y_n$ are all independent, so Y_1 and $\sum_{i=2}^n Y_i^2$ are independent. □

Expectation The above theorem tells us that $E[(\frac{n-1}{\sigma^2})S^2] = n-1$, and thus $E[S^2] = \sigma^2$

Without Normality Suppose $X_1, ..., X_n$ are i.i.d with $E[X_1] = \mu$ and $Var(X_1) = \sigma^2$, but suppose their distribution is not normal. *We still have* $E[\bar{X}_n] = \mu$, and $Var(\bar{X}_n) = \frac{\sigma^2}{n}$, and $E[S^2] = \sigma^2$. *However,* $\bar{X}_n$ is *not* necessarily normal (although it is approximately normal for large n by CLT), and the distribution of $(\frac{n-1}{\sigma^2})S^2$ is not necessarily chi-squared. *Further,* $\bar{X}_n$ and S^2 are not necessarily independent.

Student's T-Distribution

Definition Let $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi_p^2$ be independent R.V.'s, the distribution of $\frac{Z}{\sqrt{U/p}}$ is *student's t-distribution with p-degrees of freedom.* It is centered around 0.

Thrm. Let $X_1, ..., X_n \overset{i.i.d}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $n \ge 2$, then $\frac{\bar{X}_n - \mu}{\sqrt{S^2/n}} \sim t_{n-1}$.

Proof Let $Z = \frac{\bar{X}_n - \mu}{\sqrt{\sigma^2/n}}$ and $U = (n-1)S^2/\sigma^2$, by our last theorem $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi_{n-1}^2$, and they are independent. The result follows by definition since $T = \frac{Z}{\sqrt{U/(n-1)}}$ □

Lemma Let $U_n \sim \chi_n^2$ for every $n \ge 1$. Then $U_n/n \xrightarrow{P} 1$ as $\lim n \to \infty$. **Proof:** Let $Z_1, ..., Z_n \overset{i.i.d.}{\sim} \mathcal{N}(0,1)$, and let $U_n = \sum_{i=1}^n Z_i^2$. $U_n/n \xrightarrow{P} 1$ by WLLN, therefore $U_n/n \xrightarrow{D} 1$. □

Thrm. Let $T_n \sim t_n$ for every $n \ge 1$. Then $T_n \xrightarrow{D} \mathcal{N}(0,1)$ as $\lim n \to \infty$. **Proof:** Let $Z \sim \mathcal{N}(0,1)$ and $U \sim \chi_n^2$, and let $Z \perp\!\!\!\perp U$. The results follow using the Continuous Mapping Thrm., the above lemma, and Slutsky's Thrm. □.

Maximum Likelihood Estimation

Likelihood Describes the probability of observing data given certain parameter values. It is *not* a "pdf" of θ given the data x.

Thrm. Let $\hat{\theta}^{mle}$ be a maximum likelihood estimator of θ over the parameter space Θ, and let g be a function that with domain Θ and image Ξ. Then $\hat{\xi}^{mle} = g(\hat{\theta}^{mle})$ is a maximum likelihood estimator of $\xi = g(\theta)$ over the parameter space Ξ.

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown. Find the MLE of both parameters.
$L_{\boldsymbol{x}}(\mu, \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{(X_i - \mu)^2}{2\sigma^2}\right] =$
$(2\pi\sigma^2)^{-n/2} \exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2\right]$, therefore,
$\ell_{\boldsymbol{x}}(\mu, \sigma^2) = -\frac{n}{2}\ln(2\pi) - \frac{n}{2}\ln(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^n (x_i - \mu)^2$.

Differentiating w.r.t. each parameter yields:
$\frac{\partial}{\partial\mu}\ell_x(\mu,\sigma^2)=\frac{1}{\sigma^2}\sum_{i=1}^n x_i-\frac{n\mu}{\sigma^2}=\frac{n}{\sigma^2}(\bar{x_n}-\mu)$, and
$\frac{\partial}{\partial\sigma^2}\ell_{\boldsymbol{x}}(\mu,\sigma^2)=-\frac{n}{2\sigma^2}+\frac{1}{2(\sigma^2)^2}\sum_{i=1}^n(x_i-\mu)^2=\frac{1}{2(\sigma^2)^2}\sum_{i=1}^n\left[(x_i-\mu)^2-\sigma^2\right]$.
Setting both sides to zero, first note that:
$\frac{\partial}{\partial\mu}\ell_x(\mu,\sigma^2)=\frac{n}{\sigma^2}(\bar{x}-\mu)=0\implies\bar{x}=\mu$.
Substitute $\mu=\bar{x}$ in our other partial derivative and set it to 0:
$\frac{\partial}{\partial\sigma^2}\ell_{\boldsymbol{x}}(\mu,\sigma^2)=\frac{1}{2(\sigma^2)^2}\sum_{i=1}^n\left[(x_i-\mu)^2-\sigma^2\right]=0\implies\sigma^2=n^{-1}\sum_{i=1}^n(x_i-\bar{x})^2=\left(\frac{n-1}{n}S^2\right)$

Tips **1.** Check 2^{nd} derivative. **2.** Check Boundaries. **3.** Ensure estimator's max/min are within parameter space.

Bayesian Estimation

Conjugate Priors A family of distributions is called *conjugate* for a particular likelihood function if choosing a prior from that family leads to a posterior that is also from that family.

Example Let $X_1,...,X_n|\mu\overset{i.i.d.}{\sim}\mathcal{N}(\mu,\sigma^2)$, where $\mu\in\mathbb{R}$ is unknown but $\sigma^2>0$ is known. Let the prior on μ be $\mu\sim\mathcal{N}(\xi,\tau^2)$, where $\xi\in\mathbb{R}$ and $\tau^2>0$ are known. To find the posterior of μ, we use the shortcut method, ignoring anything that is not a function of μ: $L_{\boldsymbol{x}}(\mu)\pi(\mu)\propto\exp\left[-\frac{1}{2\sigma^2}\sum_{i=1}^n(x_i-\mu)^2\right]\exp\left[-\frac{(\mu-\xi)^2}{2\tau^2}\right]$
$\propto\exp\left[\frac{\mu}{\sigma^2}\sum_{i=1}^n x_i-\frac{n\mu^2}{2\sigma^2}-\frac{\mu^2}{2\tau^2}+\frac{\xi\mu}{\tau^2}\right]$
$\propto\exp\left[-\frac{(n\tau^2+\sigma^2)\mu^2}{2\sigma^2\tau^2}+\frac{(n\bar{x}\tau^2+\xi\sigma^2)\mu}{\sigma^2\tau^2}\right]$
$\propto\exp\left[-\frac{1}{2}\left(\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}\right)\left(\mu^2-2\mu\frac{n\tau^2\bar{x}+\sigma^2\xi}{n\tau^2+\sigma^2}\right)\right]$
$\propto\exp\left[-\frac{1}{2}\left(\frac{n\tau^2+\sigma^2}{\sigma^2\tau^2}\right)\left(\mu-\frac{n\tau^2\bar{x}+\sigma^2\xi}{n\tau^2+\sigma^2}\right)^2\right]$, which we recognize as another normal distribution. Thus, the posterior distribution of μ given $\boldsymbol{X}=\boldsymbol{x}$ is: $\mu|\boldsymbol{x}\sim N\left(\frac{n\tau^2\bar{x}+\sigma^2\xi}{n\tau^2+\sigma^2},\frac{\sigma^2\tau^2}{n\tau^2+\sigma^2}\right)$, which can be rewritten as $\mu|\boldsymbol{x}\sim N\left[\frac{\frac{1}{(\sigma^2/n)}}{\frac{1}{\tau^2}+\frac{1}{\sigma^2/n}}\bar{x}+\frac{\frac{1}{\tau^2}}{\frac{1}{\tau^2}+\frac{1}{\sigma^2/n}}\xi,\frac{1}{\frac{1}{\tau^2}+\frac{1}{\sigma^2}}\right]$

Estimators - Finite Sample

Bias The *bias* of an estimator $\hat{\theta}$ of a parameter θ is $Bias_\theta(\hat{\theta})=E_\theta(\hat{\theta})-\theta$. The estimator $\hat{\theta}$ is unbiased if $Bias_\theta(\hat{\theta})=0$ for all θ in the parameter space Θ.

Example Let $X_1,...,X_n$ be i.i.d. random variables such that both $\mu=E_{(\mu,\sigma^2)}(X_1)$ and $\sigma^2=Var_{(\mu,\sigma^2)}(X_1)$ are finite, and suppose $n\geq 2$. Let $\bar{X}$ and S^2 be the usual sample mean and sample variance, respectively. Then:
$E_{(\mu,\sigma^2)}(S^2)=\frac{1}{n-1}E_{(\mu,\sigma^2)}\left(\sum_{i=1}^n X_i^2-n\bar{X}^2\right)=\frac{1}{n-1}\left[n(\mu^2+\sigma^2)-n(\mu^2+\frac{\sigma^2}{n})\right]=\frac{n-1}{n-1}\sigma^2=\sigma^2$.

Example $Bias_{(\mu,\sigma^2)}\left[\frac{(n-1)S^2}{n}\right]=E_{(\mu,\sigma^2)}\left[\frac{(n-1)S^2}{n}\right]-\sigma^2=\frac{(n-1)\sigma^2}{n}-\sigma^2=-\frac{\sigma^2}{n}$, which is negative $\forall\sigma^2>0$. $\implies$ This estimator *tends* to underestimate the true value of σ^2, on average.

Variance It can also be useful to consider the variance of an estimator.

Example Suppose $X_1,...,X_n\overset{i.i.d.}{\sim}\mathcal{N}(\mu,\sigma^2)$. Then: $Var_{(\mu,\sigma^2)}(S^2)=\left(\frac{\sigma^2}{n-1}\right)^2Var_{(\mu,\sigma^2)}\left[\frac{(n-1)S^2}{\sigma^2}\right]=\left(\frac{\sigma^2}{n-1}\right)^2[2(n-1)]=\frac{2(\sigma^2)^2}{n-1}$, noting that $(n-1)S^2/\sigma^2\sim\chi^2_{n-1}$ since $X_1,...,X_n\overset{i.i.d.}{\sim}\mathcal{N}(\mu,\sigma^2)$. It follows that: $Var_{(\mu,\sigma^2)}\left[\left(\frac{n-1}{n}\right)S^2\right]=\left(\frac{n-1}{n}\right)^2Var_{(\mu,\sigma^2)}(S^2)$, which is less than the variance of S^2.

Trade-off Usually, when comparing sensible estimators, those with larger bias often have smaller variance. To get a better idea of how to compare estimators, use *Mean Squared Error*.

Mean Squared Error The M.S.E. of an estimator $\hat{\theta}$ of a parameter θ is $MSE_\theta(\hat{\theta})=E_\theta[(\hat{\theta}-\theta)^2]$.

Thrm. Let $\hat{\theta}$ be an estimator of *theta*. Then, $MSE_\theta(\hat{\theta})=[Bias_\theta(\hat{\theta})]^2+Var_\theta(\hat{\theta})$.

Proof $MSE_\theta(\hat{\theta})=E_\theta\left[(\hat{\theta}-\theta)^2\right]=\left[E_\theta(\hat{\theta}-\theta)\right]^2+Var_\theta(\hat{\theta}-\theta)=\left[Bias_\theta(\hat{\theta})\right]^2+Var_\theta(\hat{\theta})$ $\square$

Thrm Let $\gamma_w(\hat{\theta})=\int_\Theta MSE_\theta(\hat{\theta})w(\theta)\,d\theta$. Let $\hat{\theta}^B$ denote the posterior mean of θ under the prior $\pi(\theta)=w(\theta)$. Then, $\gamma_w(\hat{\theta}^B)\leq\gamma_w(\hat{\theta})$ for any other estimator $\hat{\theta}$ of θ.

Finding Unbiased Estimators No ironclad solution: **(1)** Look at $E[X]$ and $Var(X)$, play with $E[X]$, $E[X^2]$ and $E[X]^2$ to get something that looks like we're trying to estimate. **(2)** Solve for MLE. Check it's bias, adjust. **(3)** Find a function that "combines" with our pdf. E.g. $X_1,...,X_n\overset{i.i.d.}{\sim}$ Expo(λ). One attempt (which fails) is to try: e^{-cx}, via $E\left[e^{-cx}\right]$

Showing Unbiased Estimators fail to exist Estimators map observed data to estimates. Let t_i be the value our estimator takes on when we observe $x=i$. Using LOTUS, $E_\theta[\hat{\theta}]=\sum_{i=1}^n t_i\,Pr(X=i)$. Sometimes, the form of this expectation implies we can't be unbiased.

Estimators - Consistency

Definition An estimator $\hat{\theta}_n$ is a consistent estimator of a parameter θ if $\hat{\theta}_n\overset{P}{\to}\theta$ for all $\theta\in\Theta$.

Lemma Suppose $\mu=E_\mu(X_1)$ is finite, and let $\bar{X}_n$ be the usual sample mean of an i.i.d. sample $X_1,...,X_n$. If α_n is any sequence such that $\alpha_n\to 1$, then $\alpha_n\bar{X}_n$ is a consistent estimator of μ.

Thrm. If $E(X_n)\to\alpha\in\mathbb{R}$ and $Var(X_n)\to 0$, then $X_n\overset{P}{\to}\alpha$, via Chebyschev's Inequality and the definition of convergence in probability. *These conditions are sufficient, but **not** necessary!*

Corollary If $E_\theta(\hat{\theta}_n)\to\theta$ and $Var_\theta(\hat{\theta}_n)\to 0$ for all $\theta\in\Theta$, then $\hat{\theta}_n$ is a consistent estimator of θ. *These conditions are sufficient, but **not** necessary!.*

Regularity Conditions **1.** The data $\boldsymbol{X}=(X_1,...,X_n)$ is an i.i.d. sample with likelihood $L_{\boldsymbol{x}}(\theta)=\prod_{i=1}^n L_{x_i}(\theta)$ **2.** The parameter space Θ is an open subset of $\mathbb{R}$ (note that $\Theta=\mathbb{R}$ is allowed) **3.** The set $\chi=\{x_1\in\mathbb{R}:L_{x_1}(\theta)>0\}$ (called the support) does not depend on θ. **4.** If $L_{x_1}(\theta_1)=L_{x_1}(\theta_2)$ for all $x_1\in\chi$ (except possibly for some set of values with probability zero), then $\theta_1=\theta_2$. **5.** The likelihood $L_{x_1}(\theta)$ must satisfy certain smoothness conditions as a function of θ.

Thrm Let $\hat{\theta_n}$ be the MLE of θ based on the sample $\boldsymbol{X}_n=(X_1,...,X_n)$. Then under the regularity conditions above, $\hat{\theta}_n$ is a consistent estimator of θ.

Bias Vs. Consistency

Let $Y_1,...,Y_n\overset{i.i.d.}{\sim}Bern(\theta)$. Example estimators, $\hat{\theta}$, for θ:

	Consistent	Not Consistent
Unbiased	$\frac{\sum_{i=1}^n Y_i}{n}$	Y_1
Not Unbiased	$(1+\frac{1}{n})\frac{\sum_{i=1}^n Y_i}{n}$	1

Example Suppose that $\hat{\theta}$ is an unbiased estimator for θ. Is $\hat{\theta}^2$ unbiased for θ^2? **No.** Although $E_\theta[\hat{\theta}]=\theta$, $E_\theta[\hat{\theta}^2]=\left(E_\theta[\hat{\theta}]\right)^2+Var_\theta(\hat{\theta})=\theta^2+Var(\hat{\theta}^2)\geq\theta^2$, where $Var_\theta(\hat{\theta})$ non-zero unless our estimator is a constant.

Conjugate Prior Examples

Likelihood	Parameter	Conjugate Prior	Prior Hyper	Post. Hyper
Multinomial	$\boldsymbol{p}$ prob vector, k	Dirichlet	α	$\alpha+(c_1,...,c_k)$ (c_i is num. obs. in cat i)
Hypergeom. N pop. size	M (target members)	Beta-binomial	$n=N,\alpha,\beta$	$\alpha+\sum_{i=1}^n x_i,\beta+\sum_{i=1}^n N_i-\sum_{i=1}^n x_i$
Normal, known σ^2	μ	Norm	μ, σ^2	$\frac{\left(\frac{\mu_0}{\sigma_0^2}+\frac{\sum_{i=1}^n x_i}{\sigma^2}\right)}{\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)},\left(\frac{1}{\sigma_0^2}+\frac{n}{\sigma^2}\right)^{-1}$
Normal, known μ	σ^2	Inv. Gamma	α,β	$\alpha+\frac{n}{2},\ \beta+\frac{\sum_{i=1}n(x_i-\mu)^2}{2}$
Normal, known μ	σ^2	Scaled Inv. χ^2	ν,σ^2	$\nu+n,\ \frac{\nu\sigma_0^2+\sum_{i=1}^n(x_i-\mu)^2}{\nu+n}$
Uniform	$U(0,\theta)$	Pareto	x_m,k	$max\{x_1,...,x_n,x_m\},k+n$
Pareto, known min x_m	k (shape)	Gamma	α,β	$\alpha+n,\beta+\sum_{i=1}^n\ln\left(\frac{x_i}{x_m}\right)$
Weibull, known β	θ	Inv. Gamma	α,β	$\alpha+n,\beta+\sum_{i=1}^n x_i^\beta$
Inv. Gamma known α	β (inv. scale)	Gamma	α_0,β_0	$\alpha_0+n\alpha,\beta_0+\sum_{i=1}^n\frac{1}{x_1}$

Asymptotic Distribution - MLE

Score The *score* or *score-function* is simply $\ell'_{\boldsymbol{X}}(\theta)=\sum_{i=1}^n \ell'_{X_i}(\theta)$.

Information The *information* or *Fisher Information* is $I_n(\theta)=E_\theta\left[\{\ell'_{\boldsymbol{X}}(\theta)\}^2\right]$

Lemma *Under Regularity Conditions*, $E_\theta[\ell'_{\boldsymbol{X}}(\theta)]=0$, *and*

$$I_n(\theta)=\text{Var}\left[\ell'_{\boldsymbol{X}}(\theta)\right]=-E_\theta\left[\ell''_{\boldsymbol{X}}(\theta)\right]=-nE_\theta\left[\ell''_{X_1}(\theta)\right]$$

Information per Observation $I_1(\theta)=-E_\theta\left[\ell''_{X_1}(\theta)\right]$

Thrm. Let $\hat{\theta}_n$ be a maximum likelihood estimator of θ based on the sample $\boldsymbol{X}=(X_1,...,X_n)$. Then under regularity conditions,

$$\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{D}\mathcal{N}\left[0,\frac{1}{I_1(\theta)}\right]$$

Proof The basic idea is to begin with a Taylor Expansion of $\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n)$ around θ:
$\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n)=\ell'_{\boldsymbol{X}_n}(\theta)+(\hat{\theta}_n-\theta)\ell''_{\boldsymbol{X}_n}(\theta)+...$, where we ignore higher order terms based on regularity conditions. Observe that $\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n)=0$, so rearrange and multiply by $\sqrt{n}$ to get:
$\sqrt{n}(\hat{\theta}_n-\theta)\approx-\sqrt{n}\left[\frac{\ell'_{\boldsymbol{X}_n}(\theta)}{\ell''_{\boldsymbol{X}_n}(\theta)}\right]=\frac{\sqrt{n}\left[\frac{1}{n}\ell'_{\boldsymbol{X}_n}(\theta)-0\right]}{-\frac{1}{n}\ell''_{\boldsymbol{X}_n}(\theta)}$. Note that $E_\theta\left[\ell'_{\boldsymbol{X}_n}(\theta)\right]=0$ and that $\text{Var}\left[\ell'_{\boldsymbol{X}_n}(\theta)\right]=I_1(\theta)$, then by CLT: $\sqrt{n}\left[\frac{1}{n}\ell'_{\boldsymbol{X}_n}(\theta)-0\right]\xrightarrow{D}\mathcal{N}\left[0,I_1(\theta)\right]$. Also observe that the WLLN implies
$-\frac{1}{n}\ell''_{\boldsymbol{X}_n}(\theta)=-\frac{1}{n}\sum_{i=1}^n\ell''_{X_i}(\theta)\xrightarrow{P}-E_\theta\left[\ell''_{\boldsymbol{X}_1}(\theta)\right]=I_1(\theta)$
Finally, by Slutsky's Thrm., $\sqrt{n}(\hat{\theta}_n-\theta)\xrightarrow{D}\mathcal{N}\left[0,\frac{1}{I_1(\theta)}\right]$ since the asymptotic variance is $I_1(\theta)/\left[I_1(\theta)\right]^2=1/I_1(\theta)$

Observed Information Define the random variable $J_n=-\ell''_{\boldsymbol{X}_n}(\hat{\theta}_n^{mle})$. *Under regularity conditions*, $\frac{J_n}{n}$ is a consistent estimator of $I_1(\theta)$ i.e. $\frac{J_n}{n}\xrightarrow{P}I_1(\theta)$ for all $\theta\in\Theta$

Lemma Using Slutsky's and above theorem:

$$\sqrt{J_n}(\hat{\theta}_n^{mle}-\theta)=\sqrt{\frac{\frac{J_n}{n}}{I_1(\theta)}}\sqrt{nI_1(\theta)}(\hat{\theta}_n^{mle}-\theta)\xrightarrow{D}\mathcal{N}(0,1)$$

Asymptotic Efficiency

Asymptotic Variance For estimators which can be categorized by: $\sqrt{n}(\tilde{\theta}_n-\theta)\xrightarrow{D}\mathcal{N}\left[0,v(\theta)\right]$ for some function $v(\theta)$...the *asymptotic variance* of $\tilde{\theta}_n$ is given by $v(\theta)$, even though $\text{Var}(\tilde{\theta}_n)=\frac{v(\theta)}{n}$

Asymptotic Relative Efficiency If $\tilde{\theta}^{(1)}$ and $\tilde{\theta}^{(2)}$ are estimators of θ such that: $\sqrt{n}\left[\tilde{\theta}^{(1)}-\theta\right]\xrightarrow{D}\mathcal{N}\left[0,v^{(1)}(\theta)\right]$ and $\sqrt{n}\left[\tilde{\theta}^{(2)}-\theta\right]\xrightarrow{D}\mathcal{N}\left[0,v^{(2)}(\theta)\right]$, then
$ARE_\theta\left[\tilde{\theta}^{(1)},\tilde{\theta}^{(2)}\right]=\frac{1/v^{(1)}(\theta)}{1/v^{(2)}(\theta)}=\frac{v^{(2)}(\theta)}{v^{(1)}(\theta)}$

Interpretation - Sample Sizes Suppose that $ARE_\theta\left[\tilde{\theta}^{(1)},\tilde{\theta}^{(2)}\right]=3$, then the distribution of $\tilde{\theta}^{(1)}$ based on sample size n is approximately the same as the distribution of $\tilde{\theta}^{(2)}$ based on a sample of $3n$. In other words, an estimator that's 3x more efficient as another, based on ARE, needs a sample 1/3 of the size in order to achieve the same precision.

Thrm Let $\tilde{\theta}_n$ be an estimator of θ such that $\sqrt{n}(\tilde{\theta}_n-\theta)\xrightarrow{D}\mathcal{N}\left[0,v(\theta)\right]$ holds for some $v(\theta)$. Then *under regularity conditions*, $v(\theta)\geq\left[I_1(\theta)\right]^{-1}$

Asymptotic Efficiency An estimator for which $\sqrt{n}(\tilde{\theta}_n-\theta)\xrightarrow{D}\mathcal{N}\left[0,v(\theta)\right]$ holds with $v(\theta)=\left[I_1(\theta)\right]^{-1}$ is called *asymptotically efficient.*

Corollary Let $\hat{\theta}_n^{mle}$ be the MLE estimator of θ based on the sample $\boldsymbol{X}_n=(X_1,...,X_n)$. Then *under regularity conditions*, the estimator $\hat{\theta}_n^{mle}$ is asymptotically efficient.

Example - Efficiency of Bayes Estimator Let $X_1,...,X_n\overset{i.i.d.}{\sim}\text{Pois}(\lambda)$, where $\lambda>0$ is unknown. It can be shown that the posterior mean of λ under a Gamma(a,b) prior is: $\hat{\lambda}^B=\frac{a+\sum_{i=1}^n X_i}{b+n}=\left(\frac{n}{b+n}\right)\bar{X}_n+\left(\frac{b}{b+n}\right)\frac{a}{b}$. Now observe that $\sqrt{n}(\hat{\lambda}^B-\lambda)=$
$\sqrt{n}\left[\left(\frac{n}{b+n}\right)\bar{X}_n+\left(\frac{b}{b+n}\right)\frac{a}{b}\right]-\sqrt{n}\left[\left(\frac{n}{b+n}\right)\lambda+\left(\frac{b}{b+n}\right)\lambda\right]$

$$=\underbrace{\left(\frac{n}{b+n}\right)}_{\to 1}\underbrace{\sqrt{n}(\bar{X}_n-\lambda)}_{\xrightarrow{D}\mathcal{N}(0,[I_1(\theta)]^{-1})}+\underbrace{\sqrt{n}\left(\frac{b}{b+n}\right)\left(\frac{a}{b}-\lambda\right)}_{\to 0}$$

$\xrightarrow{D}\mathcal{N}\left[0,\frac{1}{I_1(\theta)}\right]$by Slutsky's Theorem. Thus $\hat{\lambda}^B$ is also asymptotically efficient.

Hypothesis Testing

Simple A hypothesis is *simple* if it fully specifies the distribution of the data (including all unknown parameter values).

Composite A hypothesis is *composite* if it is not simple.

Examples Let $X_1,...,X_n\overset{i.i.d.}{\sim}\mathcal{N}(\mu,\sigma^2)$ **1.** If $H_0:\mu=40$ and $H_1:\mu=45$ with σ^2 known. H_0 and H_1 are both simple. **2.** If $H_0:\mu=40$ and $H_1:\mu\neq40$ with σ^2 known. H_0 is simple, and H_1 composite. **3.** If $H_0:\mu=40$ and $H_1:\mu\neq40$ with σ^2 unknown. H_0 and H_1 are both composite. **4.** If $H_0:\mu\leq40$ and $H_1:\mu>40$. H_0 and H_1 are both composite. **5.** If $H_0:(\mu,\sigma^2)=(40,9)$ and $H_1:(\mu,\sigma^2)\neq(40,9)$. H_0 simple and H_1 composite.

Nested Regions Note that if $c_1>c_2$, then $R_{c_1}\subseteq R_{c_2}$.

Good Tests Mathematically, we desire that $P_\theta(\boldsymbol{X}\in R)$ tends to be higher for $\theta\in\Theta_1$ than for $\theta\in\Theta_0$. The perfect test would have $P_\theta(\boldsymbol{X}\in R)=1\,\forall\,\theta\in\Theta_1$

Type I Error A *type I error* occurs if we reject H_0 when it's actually true. i.e. if $\theta\in\Theta_0$ and $\boldsymbol{X}\in R$.

Type II Error A *type II error* occurs if we fail to reject H_0 when it's actually false. i.e. $\theta\in\Theta_1$ but $\boldsymbol{X}\notin R$.

Truth	Data	Decision	Outcome
$H_0:\theta\in\Theta_0$	$\boldsymbol{X}\notin R$	Fail to reject	Correct Decision
$H_0:\theta\in\Theta_0$	$\boldsymbol{X}\in R$	Reject H_0	Type I Error
$H_1:\theta\in\Theta_1$	$\boldsymbol{X}\notin R$	Fail to reject	Type II Error
$H_1:\theta\in\Theta_1$	$\boldsymbol{X}\in R$	Reject H_0	Correct Decision

Power Function

$$Power(\theta)=Pr_\theta(\boldsymbol{X}\in R)=\begin{cases}P_\theta(\text{type I error}) & :\theta\in\Theta_0\\ 1-P_\theta(\text{type II error}) & :\theta\in\Theta_1\end{cases}$$

Error Trade-Off If we increase c, then we tend to decrease $P_\theta(\boldsymbol{X}\in R_c)=P_\theta\left[T(\boldsymbol{X})\geq c\right]$ for all θ. This decreases the probability of a type I error but increases the probability of type II error. If we decrease c, then we tend to increase $P_\theta(\boldsymbol{X}\in R)=P_\theta\left[T(\boldsymbol{X})\geq c\right]$ for all θ. This decreases the probability of a type II error but increases the probability of a type I error.

Level of a test is any $\alpha\in\mathbb{R}$ such that $Power(\theta)\leq\alpha$ for all $\theta\in\Theta_0$. (an "upper-bound" for a type I error)

Size of a test is $sup_{\theta\in\Theta_0}Power(\theta)$. (max type I error probability)

Achieving Specified Size If the distribution of $T(\boldsymbol{X})$ is *continuous*, there exists a choice of the critical value c that achieves size α. We want to find a value $c\in R$ such that $\alpha=Pr_{\theta_0}\left[T(\boldsymbol{X})\geq c\right]=Pr_{\theta_0}\left[T(\boldsymbol{X})>c\right]=1-F_{\theta_0}^{[T(\boldsymbol{X})]}(c)$. If the distribution of $T(\boldsymbol{X})$ is *discrete*, there may not exist a c such that $Pr_{\theta_0}\left[T(\boldsymbol{X})\geq c\right]$, in which case we typically try and find a test with size less than α so it still has level α.

P-Values Suppose that we observe $\boldsymbol{X}=\boldsymbol{x}_{obs}$, then the p-value of the test with statistic $T(\boldsymbol{X})$ for the observed data is: $p(\boldsymbol{x}_{obs})=\sup_{\theta\in\Theta_0}Pr_\theta\left[T(\boldsymbol{X})\geq T(\boldsymbol{x}_{obs})\right]$.

Thrm. Let R_c be a rejection region of the form $R_c=\{\boldsymbol{x}:T(\boldsymbol{X})\geq c\}$, where c is the smallest number such that the test associated with R_c has level α. Then $\boldsymbol{x}_{obs}\in R_c\iff p(\boldsymbol{x}_{obs})\leq\alpha$.

Proof Suppose that $\boldsymbol{x}_{obs}\in R_c$. Then $T(\boldsymbol{X}_{obs})\geq c$, so $p(\boldsymbol{x}_{obs})=\sup_{\theta\in\Theta_0}Pr_\theta\left[T(\boldsymbol{X})\geq T(\boldsymbol{x}_{obs})\right]\leq\sup_{\theta\in\Theta_0}Pr_\theta\left[T(\boldsymbol{X})\geq c\right]\leq\alpha$, since the test has level α. Now suppose instead that $\boldsymbol{x}_{obs}\notin R_c$. Then $T(\boldsymbol{x}_{obs})<c$, so $p(\boldsymbol{x}_{obs})=\sup_{\theta\in\Theta_0}Pr_\theta\left[T(\boldsymbol{X})\geq T(\boldsymbol{x}_{obs})\right]>\alpha$, since otherwise c would not be the smallest number such that the test associated with R_c has level α $\square$

Corollary An equivalent way to make the final decision in a hypothesis test is to calculate the p-value $p(\boldsymbol{x}_{obs})$ for the observed data $\boldsymbol{x}_{obs}$ and reject H_0 at level α if and only if $p(\boldsymbol{x}_{obs})\leq\alpha$.

Likelihood Ratio Test

General Method Sometimes, it's not clear which test-statistic to use. The LRT is a general method based on the likelihood function, $L_{\boldsymbol{x}}(\theta)$ and the sets Θ_0 and Θ_1.

Definition Let $\Theta=\Theta_0\cup\Theta_1$. The *Likelihood Ratio Statistic* is defined as $\Lambda(\boldsymbol{X})=\frac{\sup_{\theta\in\Theta_0}L_{\boldsymbol{x}}(\theta)}{\sup_{\theta\in\Theta}L_{\boldsymbol{x}}(\theta)}$, which rejects H_0 if and only if $\Lambda(\boldsymbol{X})\leq k$, where $k\in(0,1)$ is chosen to specify the level of the test. By definition, $0\leq\Lambda(\boldsymbol{X})\leq1$.

Simple Null If **1.** The null hypothesis is simple ($H_0:\theta=\theta_0$) and **2.** The MLE $\hat{\theta}_n^{mle}$ of θ on the parameter space $\Theta=\Theta_0\cup\Theta_1$ exists, then $\Lambda(\boldsymbol{X})=\frac{L_{\boldsymbol{x}}(\theta_0)}{L_{\boldsymbol{x}}(\hat{\theta}_n^{mle})}$

Example Let $X_1,...,X_n\overset{i.i.d.}{\sim}\text{Expo}(\lambda)$, where $\lambda>0$. $H_0:\lambda=2$ and $H_1:\lambda\neq2$. Then, $L_{\boldsymbol{x}}(\lambda)=\lambda^n\exp\left(-\lambda\sum_{i=1}^n X_i\right)$. Further, $\hat{\lambda}_n^{mle}=(\bar{X}_n)^{-1}$. Then, $L_{\boldsymbol{x}}(2)=2^n\exp\left(-2\sum_{i=1}^n X_i\right)=\exp\left[-n\left(2\bar{X}_n-\log 2\right)\right]$,

and
$L_{\boldsymbol{x}}(\hat{\lambda}_n^{mle}) = \left(\frac{n}{\sum_{i=1}^n X_i}\right)^n \exp(-n) = \exp\left[-n\left(1+\log\bar{X}_n\right)\right]$.
The LRT is given by
$\Lambda(\boldsymbol{X}) = \frac{L_{\boldsymbol{x}}(2)}{L_{\boldsymbol{x}}(\hat{\lambda}_n^{mle})} = \frac{\exp\left[-n\left(2\bar{X}_n - \log 2\right)\right]}{\exp\left[-n\left(1+\log\bar{X}_n\right)\right]} =$
$\exp\left[n\left(1+\log 2+\log\bar{X}_n - 2\bar{X}_n\right)\right] = \left[2\bar{X}_n \exp\left(1-2\bar{X}_n\right)\right]^n$.
Ultimately, LRT rejects H_0 if and only if
$\left[2\bar{X}_n \exp\left(1-2\bar{X}_n\right)\right]^n \leq k$. Equivalently, reject if
$\bar{X}_n \exp\left(-2\bar{X}_n\right) \leq (2e)^{-1}k^{1/n}$

Composite Null In this case, finding the numerator of $\Lambda(\boldsymbol{X})$ typically requires first maximizing the likelihood function subject to the constraints of H_0, then evaluating the likelihood at this point.

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown. $H_0 : \mu = \mu_0$, $H_1 : \mu \neq \mu_0$ for some $\mu_0 \in \mathbb{R}$. The numerator of $\Lambda(\boldsymbol{X})$ is given by $\sup_{\sigma^2>0} L_{\boldsymbol{x}}(\mu_0, \sigma^2)$. Observe that
$\frac{\partial}{\partial\sigma^2}\ell_{\boldsymbol{x}}(\mu_0, \sigma^2) = -\frac{n}{n\sigma^2} + \frac{1}{2(\sigma^2)^2}\sum_{i=1}^n (X_i - \mu_0)^2 = 0 \iff$
$\tilde{\sigma}_0^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2$, since this value is indeed a maximum. Recall the unconstrained MLE of μ and σ^2 are given by: $\hat{\mu} = \bar{X}_n$ and $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \bar{X}_n)^2$. The LRT is given by:
$\Lambda(\boldsymbol{X}) = \frac{(2\pi\tilde{\sigma}_0^2)^{-n/2}\exp\left[-(2\tilde{\sigma}_0^2)^{-1}\sum_{i=1}^n (X_i-\mu_0)^2\right]}{(2\pi\hat{\sigma}^2)^{-n/2}\exp\left[-(2\hat{\sigma}^2)^{-1}\sum_{i=1}^n (X_i-\hat{\mu})^2\right]} =$
$\frac{(\tilde{\sigma}_0^2)^{-n/2}\exp[-n/2]}{(\hat{\sigma}^2)^{-n/2}\exp[-n/2]} = \left(\frac{\hat{\sigma}^2}{\tilde{\sigma}_0^2}\right)^{n/2} = \left[\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2}{\sum_{i=1}^n (X_i - \mu_0)^2}\right]^{n/2}$.
Observe that
$\sum_{i=1}^n (X_i - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X}_n + \bar{X}_n - \mu_0)^2 = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu_0)^2 + 2(\bar{X}_n - \mu_0)\sum_{i=1}^n (X_i - \bar{X}_n) = \sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu_0)^2$. Then,
$\Lambda(\boldsymbol{X}) = \left[\frac{\sum_{i=1}^n (X_i - \bar{X}_n)^2 + n(\bar{X}_n - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X}_n)^2}\right]^{-n/2} =$
$\left[1 + \frac{n(\bar{X}_n - \mu_0)^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right]^{-n/2} = \left[1 + \frac{(\bar{X}_n - \mu_0)^2}{\hat{\sigma}^2}\right]^{-n/2} =$
$\left[1 + \frac{[T(\boldsymbol{X})]^2}{n-1}\right]^{-n/2}$, where $T(\boldsymbol{X}) = \frac{|\bar{X}_n - \mu_0|}{\sqrt{\hat{\sigma}^2/(n-1)}} = \frac{|\bar{X}_n - \mu_0|}{\sqrt{S^2/n}}$
where $S^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X}_n)^2$ is the unbiased sample variance. If H_0 true, then $T(\boldsymbol{X})$ is the distribution of the absolute value of a Student's T random variable.

Wald Test

Background Under *suitable regularity conditions*, we know the asymptotic distribution of MLE's is normal:
$\sqrt{I_n(\hat{\theta}_n^{mle})}(\hat{\theta}_n^{mle} - \theta) \overset{D}{\to} \mathcal{N}(0,1)$ and also that
$\sqrt{J_n}(\hat{\theta}_n^{mle} - \theta) \overset{D}{\to} \mathcal{N}(0,1)$, where $J_n = -\ell''_{\boldsymbol{X}_n}(\hat{\theta}_n^{mle})$

Definition Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, with size α, by rejecting H_0 if and only if $\sqrt{I_n(\hat{\theta}_n^{mle})}|\hat{\theta}_n^{mle} - \theta_0| \geq c$, where c is the number such that $Pr(|Z| \geq c) = \alpha$ for a standard normal RV Z. Alternatively, reject $H_0 \iff$
$\sqrt{J_n}|\hat{\theta}_n^{mle} - \theta_0| \geq c$.

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \text{Expo}(\lambda)$, where $\lambda > 0$. Test $H_0 : \lambda = 2$ against $H_1 : \lambda \neq 2$. Recall that $\hat{\lambda}_n^{mle} = (\bar{X}_n)^{-1}$. Note that $\ell''_{\boldsymbol{X}_n}(\lambda) = \frac{\partial^2}{\partial\lambda^2}\left(n\log\lambda - \lambda\sum_{i=1}^n X_i\right) = -\frac{n}{\lambda^2}$. So, $I_n(\lambda) = -E_\lambda[\ell''_{\boldsymbol{X}_n}(\lambda)] = n/\lambda^2$, and hence
$I_n(\hat{\lambda}_n^{mle}) = \frac{n}{(\hat{\lambda}_n^{mle})^2}$. Similarly,
$J_n = -\ell_{\boldsymbol{X}_n}(\hat{\lambda}_n^{mle}) = \frac{n}{(\hat{\lambda}_n^{mle})^2}$. The Wald Tests for either form are identical, *in this case*:
$\sqrt{I_n(\hat{\lambda}_n^{mle})}|\hat{\lambda}_n^{mle} - 2| = \sqrt{J_n}|\hat{\lambda}_n^{mle} - 2| =$
$\sqrt{\frac{n}{(\hat{\lambda}_n^{mle})^2}}|\hat{\lambda}_n^{mle} - 2| = \sqrt{n}|1 - \frac{2}{\hat{\lambda}_n^{mle}}| = \sqrt{n}|1 - 2\bar{X}_n|$

Score Test

Background Recall that *under regularity conditions*,
$\sqrt{n}\left[\frac{1}{n}\ell'_{\boldsymbol{X}_n}(\theta) - 0\right] = \frac{1}{\sqrt{n}}\ell'_{\boldsymbol{X}_n}(\theta) \overset{D}{\to} \mathcal{N}[0, I_1(\theta)]$. It follows that $\frac{1}{\sqrt{nI_1(\theta)}}\ell'_{\boldsymbol{X}_n}(\theta) = \frac{1}{\sqrt{I_n(\theta)}}\ell'_{\boldsymbol{X}_n}(\theta) \overset{D}{\to} N(0,1)$

Definition Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, with approximate size α. Reject $H_0 \iff \frac{1}{\sqrt{I_n(\theta_0)}}|\ell'_{\boldsymbol{X}_n}(\theta_0)| \geq c$, where c is the number such that $Pr(|Z| \geq c) = \alpha$ for a standard normal RV Z.

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \text{Expo}(\lambda)$, where $\lambda > 0$. Test $H_0 : \lambda = 2$ against $H_1 : \lambda \neq 2$. The score function is:
$\ell'_{\boldsymbol{X}_n}(\lambda) = \frac{\partial}{\partial\lambda}\left(n\log\lambda - \lambda\sum_{i=1}^n X_i\right) = \frac{n}{\lambda} - \sum_{i=1}^n X_i =$
$n\left(\frac{1}{\lambda} - \bar{X}_n\right)$, where $\bar{X}_n = n^{-1}\sum_{i=1}^n X_i$. From previous example, $I_n(\lambda) = n/(\lambda)^2$. Then, the score test statistic is given by:
$\frac{1}{\sqrt{I_n(2)}}|\ell'_{\boldsymbol{X}_n}(2)| = \frac{1}{\sqrt{n/4}}\left|n\left(\frac{1}{2} - \bar{X}_n\right)\right| = \sqrt{n}|1 - 2\bar{X}_n|$. The score test rejects H_0 if and only if the statistic is at least as large as some value, c, determined by the size.

Asymptotic Likelihood Ratio Tests

Thrm. Let $\Lambda(\boldsymbol{X}_n)$ be the likelihood ratio test for testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, based on sample $\boldsymbol{X}_n$. Then *under regularity conditions*: $-2\log\Lambda(\boldsymbol{X}_n) \overset{D}{\to} \chi_1^2$, if $\theta = \theta_0$.

Proof Let $\hat{\theta}_n^{mle}$ denote the MLE of θ. A Taylor Expansion of $\ell_{\boldsymbol{X}_n}(\theta_0)$ around $\ell_{\boldsymbol{X}_n}(\hat{\theta}_n^{mle})$:
$$\ell_{\boldsymbol{X}_n}(\theta_0) = \ell_{\boldsymbol{X}_n}(\hat{\theta}_n) + \ell'_{\boldsymbol{X}_n}(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n) + \frac{1}{2}\ell''_{\boldsymbol{X}_n}(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)^2 + ...$$
$$= \ell_{\boldsymbol{X}_n}(\hat{\theta}_n) + \frac{1}{2}\ell''_{\boldsymbol{X}_n}(\theta_0 - \hat{\theta}_n)^2 + ...$$
since $\ell'_{\boldsymbol{X}_n}(\hat{\theta}_n) = 0$. Further, the regularity conditions allow us to ignore higher-order terms. Now, observe that
$-2\log\Lambda(\boldsymbol{X}_n) = -2\log\left[\frac{L_{\boldsymbol{X}_n}(\theta_0)}{L_{\boldsymbol{X}_n}(\hat{\theta}_n)}\right] =$
$-2\left[\ell_{\boldsymbol{X}_n}(\theta_0) - \ell_{\boldsymbol{X}_n}(\hat{\theta}_n)\right] \approx -\ell''_{\boldsymbol{X}_n}(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)^2$ by the Taylor Expansion. Then, $-2\log\Lambda(\boldsymbol{X}_n) \approx$
$-\ell''_{\boldsymbol{X}_n}(\hat{\theta}_n)(\theta_0 - \hat{\theta}_n)^2 = J_n(\hat{\theta}_n - \theta_0)^2 = \left[\sqrt{J_n}(\hat{\theta}_n - \theta_0)\right]^2$. If the true value of $\theta = \theta_0$, then
$\sqrt{J_n}|\hat{\theta}_n - \theta| \overset{D}{\to} \mathcal{N}(0,1) \implies \left[\sqrt{J_n}|\hat{\theta}_n - \theta|\right]^2 \overset{D}{\to} \chi_1^2$, using continuous mapping thrm. for convergence in distribution.

Rejection Region A test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ with approximate size α is to reject $H_0 \iff -2\log\Lambda\boldsymbol{X}_n \geq C$, where C is the number such that $Pr(W \geq C) = \alpha$ for a χ_1^2 R.V. W, or equivalently, the number such that $Pr(|Z| \geq \sqrt{C}) = \alpha$ for a $\mathcal{N}(0,1)$ R.V. Z.

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \text{Expo}(\lambda)$, where $\lambda > 0$. Test $H_0 : \lambda = 2$ against $H_1 : \lambda \neq 2$. Recall the LRT is:
$\Lambda(\boldsymbol{X}_n) = [2\bar{X}_n \exp(1 - 2\bar{X}_n)]^n$. Note that
$-2\log\Lambda(\boldsymbol{X}_n) = -2n[1 + \log(2\bar{X}_n) - 2\bar{X}_n]$. To obtain LRT with size α, reject $H_0 \iff$ test statistic is at least as large as some critical value C. To obtain size $\alpha = 0.05$, take $\sqrt{C} \approx 1.96$, hence $C \approx 3.84$.

Summary of Asymptotic Tests

Similarities The Wald Test, Score Test, and LRT provide different ways to construct hypothesis tests of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, with approximate size α for large n.

Basis: **Wald** is based on the difference between θ_0 and $\hat{\theta}_n^{mle}$. **Score** is based on the difference between the slope of the log-likelihood at θ_0 against $\hat{\theta}_n^{mle}$. **LRT** is based on the difference between the likelihood at θ_0 against $\hat{\theta}_n$.

Computation: **Wald**, when based on J_n, only requires the behavior of the log-likelihood at and around it's global max, $\hat{\theta}_n^{mle}$. **Score** only involves the behavior of the log-likelihood at and around θ_0. **LRT** involves behavior of likelihood at both θ_0 and $\hat{\theta}_n^{mle}$.

Reparametrization Score and LRT invariant, but Wald isn't.

Confidence Intervals

Definition A *Confidence Level* of a confidence set $C(\boldsymbol{X})$ for a parameter $\theta \in \Theta$ is a number $\gamma \geq 0$ such that
$Pr_\theta\left[\theta \in C(\boldsymbol{X})\right] \geq \gamma$ for all $\theta \in \Theta$.

Thrm. For every $\theta_0 \in \Theta$, let R_{θ_0} be the rejection region of a hypothesis test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$, with level α. Then $C(\boldsymbol{X}) = \{\theta_0 \in \Theta : \boldsymbol{X} \notin R_{\theta_0}\}$ is a confidence set for θ with confidence level $1 - \alpha$.

Proof For every $\theta \in \Theta$, $Pr_\theta\left[\theta \in C(\boldsymbol{X})\right] = Pr_\theta(X \notin R_\theta) = 1 - Pr_\theta(\boldsymbol{X} \in R_\theta) \geq 1 - alpha$. $\square$

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, where $\mu \in \mathbb{R}$ and $\sigma^2 > 0$ are both unknown. Let $\alpha \in (0,1)$. Begin by finding a test of $H_0 : \mu = \mu_0$ against $H_1 : \mu \neq \mu_0$ with level α. The LRT of these hypotheses is to reject $H_0 \iff \frac{|\bar{X}_n - \mu_0|}{\sqrt{S^2/n}} \geq c$. Then a confidence set for μ with confidence level $1 - \alpha$ is the set of all $\mu_0 \in \mathbb{R}$ such that $\frac{|\bar{X}_n - \mu_0|}{\sqrt{S^2/n}} < c$

Wald Confidence Interval The simplest asymptotic CI. Test $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ reject
$H_0 \iff \sqrt{I_n(\hat{\theta}_n^{mle})}|\hat{\theta}_n^{mle} - \theta_0| \geq c$, or alternatively
$\sqrt{J_n(\hat{\theta}_n^{mle})}|\hat{\theta}_n^{mle} - \theta_0| \geq c$, where c is the number such that $Pr(|Z| \geq c) = \alpha$ for $Z \sim \mathcal{N}(0,1)$. Our test fails to reject
$H_0 \iff \left\{\theta_0 \in \Theta : \hat{\theta}_n - \frac{c}{\sqrt{J_n(\hat{\theta})}} < \theta_0 < \hat{\theta}_n + \frac{c}{\sqrt{J_n(\hat{\theta})}}\right\}$

Example Let $X_1, ..., X_n \overset{i.i.d.}{\sim} (\lambda)$, where $\lambda > 0$ unknown. Let $\alpha \in (0,1)$. We found earlier that both versions of Wald reject $H_0 \iff \sqrt{\frac{n}{\hat{\lambda}_n^2}}|\hat{\lambda}_n - \lambda_0| \geq c$. The Wald CI with approximate confidence level $1 - \alpha$ is the set:
$\left\{\lambda_0 > 0 : \hat{\lambda}_n - c\sqrt{\frac{\hat{\lambda}_n^2}{n}} < \lambda_0 < \hat{\lambda}_n + c\sqrt{\frac{\hat{\lambda}_n^2}{n}}\right\}$. Note the restriction that $\lambda_0 > 0$ ensures the CI doesn't spill over the parameter space.

Score Confidence Interval A score test of $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$ rejects $H_0 \iff \frac{1}{\sqrt{I_n(\theta_0)}}|\ell'_{\boldsymbol{X}_n}(\theta_0)| \geq c$. The test fails to reject $H_0 \iff \left\{\theta_0 \in \Theta : \frac{1}{\sqrt{I_n(\theta_0)}}|\ell'_{\boldsymbol{X}_n}(\theta_0)| < c\right\}$.
This is a Score Confidence Set.