STA261 Review

Max Chen

April 2018

Sincere thanks to Alex Stringer, the professor of the course. This note is taken during and after his lecture, based on the lecture materials. Personal understanding has been added. This note should not be used for any purpose other than study and learn.

1 Week 1: Review of STA257 Convergence of RVs

1. **Random Vairable** is a function from a sample space Ω to (a subset of) R.

2. **Support of X** is the subset of R to which X maps to.

3. **expected value = expectation = mean** is the single real number that is "closest" to X in Euclidean distance.

E(aX+b) = aE(X)+b

E(g(X)) = g(E(X)) iff g is linear

4. **Standard Deviation** is the Euclidean distance from the random variable to its mean.

$Var(X) = SD(x)^2 = E(X^2) - E(X)^2$

5. **Moment-Generating Function** is $M_X(t) = E(e^{tX})$

• compute moments E(X^k) = $M_X^{(k)}(0)$

• Two RV have the same distribution iff X $=^d = Y \iff M_X(t) = M_Y(t)$

6. **Chebyshev**: P($\|X - E(X)\| >$ t) $\leq \frac{Var(X)}{t^2}$, $\forall$ t>0

Markov: X $\geq$ 0 with probability 1, and E(X) exists, then P(X $\geq t) \leq \frac{E(X)}{t}$, $\forall$ t > 0

7. **Converges in probability**: sequence Z_n converges in probability to μ if $\forall \epsilon > 0$, $\lim_{n\to\infty}(P(\|Z_n - \mu\|) > \epsilon)$ = 0, denote Z$_n \xrightarrow{p} \mu$.

Thm: Suppose Z_n is a sequence of RV with E(Z$_n$) = μ and $\lim_{n\to\infty} Var(Z_n) = 0$, then Z$_n \xrightarrow{p} \mu$.

8. LLN: Suppose X_n is a sequence of indep RV with E(X$_i$) = μ and Var(X$_i$) = $\sigma^2. Let \bar{X_n} = \frac{\sum_{i=1}^n X_i}{n}, then \bar{X_n} \xrightarrow{p} \mu$.

$\star$ average converges to mean, for large samples, i.e. n $\to \infty \implies \bar{X_n} \xrightarrow{p} \mu$.

$\star$ this also says Var($\bar{X_n}$) $\to 0$ as n $\to \infty$.

9. **Converges in Distribution**: sequence $X_n \xrightarrow{d}$ X if $\lim_n \to \infty F_n(x) = F_X(x), \forall$ x at which these distributions fcns are continuous.

Also, $\lim_n \to \infty M_n(t) = M_X(t)$ $\forall$ t $\implies X_n \xrightarrow{d} X$

$\star$ X_n and X has the same probability distribution fcn doesn't mean they are equal $\forall$ n.

10. Let c $\in$ R, then $X_n \xrightarrow{p} c \implies X_n \xrightarrow{d} c$.

Thm: Let X be "degenerated" RV with Var(X) = 0, so that P(X = c) = 1, then $X_n \xrightarrow{d} c \implies X_n \xrightarrow{p} c$

11. CLT: Let X_n be a sequence of **independent** RVs, $E(X_i) = 0$ and Var(X_i) $= \sigma^2 . Let S_n = \sum_{i=1}^n X_i . Then$ $\frac{S_n}{\sigma\sqrt{n}} \xrightarrow{d} N(0,1)$

Or: $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \xrightarrow{d} N(0,1)$

2 Week 2: Intro to Estimation Theory: Consistency MoM

1. **Family**: a set of distributions is a "family" if they have the same functional form, but are specified only up to an unknown parameter.

2. **Parameter** θ: a fixed, constant element of the vector space R^d. If d > 1, then θ is a vector.

$\hat{\theta}$**, Estimator of** θ: a function that estimates the parameter θ.

- Estimators are RVs because they are functions of RVs.

- The probability distribution of an estimator is sometimes referred to as its **sampling distribution**.

$\hat{\theta}$**, Estimate of** θ: an actual number, by plugging a real dataset into the estimator .

3. **Consistency**: $\hat{\theta}$ is consistent for θ if $\hat{\theta} \xrightarrow{p} \theta$.

As we get more data, we should be able to get close as we want to the parameter we are estimating, with as high a probability as we want.

To prove $\hat{\theta} = (\hat{\theta_1}, ..., \hat{}_d)$ is consistent for $\theta = (\theta_1, ..., \theta_d)$, just prove $\hat{\theta_k} \xrightarrow{p} \theta_k$, $\forall k = 1...d$

4. LLN $\implies \frac{\sum_{i=1}^n X_i^k}{n} \xrightarrow{p} E(X^k) \implies \bar{X^k} \xrightarrow{p} \mu^k$

$\star$Due to continuity.(Slutsky Lemma)

Application: $s^2 = \frac{\sum_{i=1}^n (X_i-\mu)^2}{n}$ or $s^2 = \frac{\sum_{i=1}^n (X_i-\bar{X})^2}{n-1}$ is a consistent estimator of σ^2

5. **Method of Moments**

Algorithm: Let $X_i \sim F_\theta$ independently, $\theta = (\theta_1, ...\theta_d)$.

- Find expressions for the first d population moments in terms of θ_1,...,θ_d,

$$E(X) = g_1(\theta_1, ..., \theta_d)$$
$$E(X^2) = g_2(\theta_1, ..., \theta_d)$$
$$...$$
$$E(X^d) = g_d(\theta_1, ..., \theta_d)$$

- Solve for $\theta_1, ..., \theta_d$.
- Apply LLN.
  The resulting estimators are **consistent, continuous and invertable**.

Ex: Let $X_i \sim$ Unif(a,b), find a MoM estimator for $\theta = (a, b)$

3 Week 3: Sufficiency Likelihood

1. **Sufficiency**: An estimator $\hat{\theta}(X_1, ..., X_n)$ is "sufficient" for the parameter θ if the conditional distribution of the sample $X_1, ..., X_n$ given $\hat{\theta} = t$ does not depend on $\theta, \forall$ t.

$\star$ Our estimator should be a *summary* of the full sample, we should make the same conclusion regarding θ.

2. **Statistics**: any function that takes in data and returns a (possibly lower-dim) summary.

A sufficient estimator is a **sufficient statistics**.

Notation: $\mathbf{X} = (X_1, ..., X_n)$ is RV, $\mathbf{x} = (x_1, ..., x_n)$ is **realization** of RV, i.e. a datapoint.

3. **Factorization Thm**: $\hat{\theta}$ is sufficient for θ iff the joint density of $X_1, ..., X_n$ can be factorized as

$$f_{\mathbf{X}}(x_1, ..., x_n) = g(\hat{\theta}, \theta) \times h(\mathbf{x})$$

Ex: Any one-to-one functions of a sufficient statistics is sufficient.

4. **Rao-Blackwell Thm**: Let $\hat{\theta}$ be any estimator of θ, $E(\hat{\theta}^2 < \infty)$. Let T be any sufficient statistics (for θ), and let $\tilde{\theta} = E(\hat{\theta}|T)$. Then

$$Var(\tilde{\theta}) \leq Var(\hat{\theta})$$

, equality holds when $\hat{\theta} = \tilde{\theta}$.

$\star$ Sufficient statistics(estimators) has smaller variance. For all sufficient statistics, some of them could have smaller variance than others.

5. **Likelihood function** is the joint distribution of the data, treated as a function of the parameters:

$$L(\theta|\mathbf{x}) = f(\mathbf{x}|\theta) = \Pi_{i=1}^{n} f_{x_i}(x_i|\theta)$$

6. **Maximum Likelihood Estimator(MLE) log-likelihood**

values of θ that give a higher $L(\theta)$ are more likely to have generated the observed data.

$\star$ maximized with respect to σ^2

$\star$ precision $= \frac{1}{\sigma^2}$

Ex: Let X_i ~Unif(0,b), find MLE of b. (Hint: $L(b) = \Pi_{i=1}^{n} \frac{1}{b} \times I(x_i \leq b) \implies \hat{b} = max(x_i)$, cannot be calculated with calculus.)

$\star$ MLE is **Consistent, Sufficient, asymptotically Unbiased and asymptotically efficient**.

4 Week 4: Likelihood inference

1. **Curvature**: $|\frac{\partial^2 f(x)}{\partial x^2}|$

$\star$ likelihood function defines which values of θ are plausible given the observed data. *Peaked* likelihood $\implies$ narrow range of plausible values for θ. *Flat* likelihood $\implies$ wide range of plausible values for θ.

2. **Score Vector** (Score function, score statistics): $S(\theta) = \frac{\partial \ell}{\partial \theta}$

$\star$ $S(\theta) = 0 \implies$ MLE.

⋆ Parameter space Ω is the set of all values that θ can take.

3. Regularity Conditions:

- true parameter in the interior of the parameter space, i.e.$\theta_0 \in \Omega_0$

- support of the distribution of $\mathbf{X}$ doesn't depend on θ.

- log-likelihood is of class C^3.

4. $E(S\theta_0) = 0$

$Var(s_i(\theta_0)) = E(s_i(\theta_0)^2) = -E(\frac{\partial^2 \log f(x_i|\theta_0)}{\partial\theta^2})$

5. **Fisher Information**: expected value of negative value of the second derivative of the log-likelihood function

$I_i(\theta) = Var(s_i(\theta))$ (of a data point)

$I_i(\theta_0) = -E(\frac{\partial^2 \ell(\theta|x_i)}{\partial\theta^2}) \mid_{\theta=\theta_0}$

$I(\theta|\mathbf{x}) = \sum_{i=1}^{n} I_i(\theta)(= nI_0(\theta)$ if IID)

6. **Observed Information**: the negative value of the second derivative of the log-likelihood function.

$J(\theta) = -\sum_{i=1}^{n} \frac{\partial^2 \ell(\theta|x_i)}{\partial\theta^2}$

$\frac{J(\theta)}{n} = -\frac{1}{n}\sum_{i=1}^{n} \frac{\partial^2 \ell(\theta|x_i)}{\partial\theta^2} \xrightarrow{p} I_0(\theta)$ (consistent estimator of $I_0(\theta)$)

⋆ Fisher info is the expected value of observed info.

⋆ in multiparameter case, also need the cross second partials.

7. Summary

- log-likelihood
- first derivative $\implies$ score vector
- second derivative $\implies$ observed info
- expectation of second derivative $\implies$ fisher info

- $E(S(\theta_0)) = 0$ and $Var(S(\theta_0)) = I(\theta_0)$

  By CLT:

- $\frac{S(\theta_0)}{\sqrt{I(\theta_0)}} = \frac{\sum_{i=1}^{n} S_i(\theta_0)}{\sqrt{nI_i(\theta_0)}} \xrightarrow{d} N(0,1)$

- $\frac{S(\theta_0)}{\sqrt{J(\theta_0)}} \xrightarrow{d} N(0,1)$

- $\sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,1)$

- "large sample distribution": MLE is approximately normally distributed with mean equal to the true value θ_0 and variance equal to the inverse of Fisher info $\frac{1}{I(\theta_0)}$ (can plug the estimator $\hat{\theta}$ for θ_0 due to consistency).

- *Asymptotic Covariance Matrix* is given by the inverse of the Information Matrix

5 Week 5: Unbiasedness Efficiency

1. **Bias**: $bias(\hat{\theta}) = E(\hat{\theta} - \theta) = E(\hat{\theta}) - \theta$

$\star$ the degree by which we expect $\hat{\theta}$ to differ from θ

$\hat{\theta}$ is **Unbiased** if $E(\hat{\theta}) = \theta \iff bias(\hat{\theta}) = 0$

Ex: $X_i \sim Exp(\beta)$, with f(x) = $\beta e^(-\beta x)$. $\hat{\beta} = \frac{1}{\bar{X}}$ is unbiased for β. (Hint: MLE is **asymptotically unbiased** because CLT $\implies E(\hat{\theta} - \theta) \to 0$)

2. Cramer-Rao Lower Bound Thm: Suppose $\hat{\theta}$ is any unbiased estimator for θ. Then

$$Var(\hat{\theta}) \geq \frac{1}{nI_0(\theta_0)}$$

, where I_0 is the Fisher Info for a single data point

3. **Efficiency**: $\hat{\theta}$ is **efficient** if it attains the Cramer-Rao Lower Bound, i.e. $Var(\hat{\theta}) = \frac{1}{nI_0(\theta_0)} = \frac{1}{I(\theta_0)}$

$\star$ MLE is *asymptotically efficient*.

6 Week 7: Confidence Intervals(CI) Hypothesis Testing I

1. **A range of plausible values**: a range of values that "could plausibly have generated the data we observed".

2. **Pivot**: a pivot for parameter θ is a RV that depends on the unknown parameter θ_0, but has a known distribution that does not depend on θ_0.
e.g. $Z = \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} \sim N(0,1)$ and $\frac{ns^2}{\sigma_0^2} \sim \chi^2_n$

3. $1-\alpha$ **CI for** μ: an interval C(X) = (L(X), U(X)) s.t. $P(L(X) \leq \mu_0 \leq U(X)) = 1-\alpha$, for some $1 < \alpha < 0.5$

$\star$ "the probability that the interval contains μ_0 is $1-\alpha$." (interval is random, μ_0 isn't.)

$\star$ σ^2 known, use $\frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$

$\star$ σ^2 unknown, use $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \implies \frac{\bar{X}-\mu}{s/\sqrt{n}} \sim t_{n-1}$
$\star$ μ known, use $s^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu)^2 \implies \frac{ns^2}{\sigma^2} \sim \chi^2_n$ (Narrower CI)

$\star$ μ unknown, use $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2 \implies \frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{n-1}$

4. **Hypothesis Test**: Null hypothesis ($H_0 : \mu = \mu_0$), alternative hypothesis H_1.

A hypothesis is **Simple** if a single value in the parameter space; **Composite** if contains more than one values.

Ex: Dimension of parameter space: $\theta = 2, \theta_0, \{\theta_i\}_{i=1}^n$?
$\star$ **Never** "accpet" the null. We say "Failed to provide sufficient evidence against the null."

5. **Type I Error**: reject the null when it's true. (worse)

P(type I error) = P(reject $H_0|H_0$ true) = α (**Significant Level**)

Type II Error: fail to reject the null when it's false.

P(type II error) = P(fail to reject $H_0|H_0$ false) = β

6. **Test Statistics**: T(X)

We choose T(X) s.t.:

- has a known distribution if H_0 is true
- depends on the data through an estimator of some kind
- $P(T(X) \in R_\alpha(T) \mid H_0 true) = \alpha$ (tractable)

7. **Critical Region**: $R_\alpha(T)$ reject H_0 if $T(X) \in R_\alpha(T)$

8. **P value**(p_0): the probability of observing a test statistics with euqal or greater evidence against H_0

$p_0 = P(T(X) > |t(x)|)$

Proposition: when H_0 is true, $p_0 \sim Unif(0,1) \implies P(p_0 < \alpha) = \alpha \implies p_0 < \alpha \implies$ reject H_0

9.Unknown Variance: replace σ with a consistent estimator, $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2} \implies T(X) = \frac{\bar{X}-\mu_0}{s/\sqrt{n}}$ (Student's Statistics)

10. Distribution of Sample Variance:

- if μ known, $\sum_{i=1}^n (\frac{X_i - \mu_0}{\sigma_0})^2 = \frac{n\hat{\sigma}^2}{\sigma_0^2} \sim \chi_n^2$,where $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^n (X_i - \mu_0)^2$
- if μ unknown, $\frac{(n-1)s^2}{\sigma^2} \sim \chi_{n-1}^2$,where $s^2 = \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$

$\star$ if Z_i, i=1...n is an IID sample from a N(0,1), then $S = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$

11. **Joint Normality** ?

12. **t-distribution**: Let $Z \sim N(0,1), U \sim \chi_\nu^2$ and $Z \perp U$. Then the Student's t-distribution with ν degrees of freedom is: $T = \frac{Z}{\sqrt{U/\nu}} \sim t_\nu$

Corollary: $T = \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \sim t_{n-1}$ (proof: $\frac{\bar{X}-\mu_0}{s/\sqrt{n}} = (\frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}) \times (\frac{s^2}{\sigma^2})^{-1/2}$)

$\star$ t-distribution is symmetric, f(t) = f(-t)

$\star$ E(T) = 0, Var(T) = $\frac{\nu}{\nu-2}$, for $\nu > 2$

$\star$ as $\nu \to \infty$, $T \xrightarrow{d} Z$

$\star$ CI: $(\bar{X} - \frac{s}{\sqrt{n}} t_{n-1,\alpha/2}, \bar{X} + \frac{s}{\sqrt{n}} t_{n-1,\alpha/2})$

7 Week 8: CI Hypothesis Testing II

1. large sample $1-\alpha$ CI for θ: $(\hat{\theta} - \frac{1}{\sqrt{I(\hat{\theta})}} z_{\frac{1-\alpha}{2}}, \hat{\theta} + \frac{1}{\sqrt{I(\hat{\theta})}} z_{\frac{1-\alpha}{2}})$, since $\sqrt{I(\theta_0)}(\hat{\theta} - \theta_0) \xrightarrow{d} N(0,1)$

2. Monotone transformation: $1-\alpha$ CI for θ = (L, U)

- If g() is monotonic increasing, then $1-\alpha$ CI for $g(\theta)$ = (g(L), g(U))
- If g() is monotonic decreasing, then $1-\alpha$ CI for $g(\theta)$ = (g(U), g(l))

3. Two-sample Problems: "Does the mean measurement differ between group A and group B?"
$\star$ mind the degrees of freedom when unknown variance.
(degree of freedom = number of parameter to estimate under alternative hypothesis - number of parameter to estimate under null hypothesis)

⋆ Two groups do not have to be of the same size, but have to assume they have the same variance.

4. Paired Sample: Same group measured before, after test.

⋆ CI of paired sample is narrower than two-sample, and we only have to sample half of the data compared to the two-sample problem.

⋆ Pooled Variance:

$$s_X^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \tag{1}$$

$$s_Y^2 = \frac{1}{m-1}\sum_{i=1}^{m}(Y_i - \bar{Y})^2 \tag{2}$$

$$\implies \hat{\sigma}^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2 + \sum_{i=1}^{m}(Y_i - \bar{Y})^2}{n+m-2} \tag{3}$$

$$\tag{4}$$

8 Week 9: Likelihood Ratio Test

1. **Likelihood Ratio**: $\Lambda = \frac{L(\mu_0)}{L(\mu_1)}$

⋆The value with higher likelihood is better supported by the data, i.e. $\Lambda > 1 \implies \mu_0$ better, $\Lambda < 1 \implies \mu_1$ better.

2. **Likelihood Ratio Test**:for testing $H_0 : \theta \in \Omega_0$ against $H_1 : \theta \in \Omega_1$ is $\Lambda = \frac{sup_{\theta\in\Omega_0} L(\mu_0)}{sup_{\theta\in\Omega_1} L(\mu_1)}$

⋆ Small $\Lambda \implies H_1$ is better supported by the data. And we reject H_0 if Λ is "small enough".

⋆ $\Lambda = \frac{sup_{\theta\in\Omega_0} L(\mu_0)}{sup_{\theta\in\Omega_1} L(\mu_1)} = \frac{sup_{\theta\in\Omega_0} L(\mu_0)}{L(\hat{\theta})}$, $\hat{\theta}$ is the MLE.

⋆ The closer $\hat{\theta}$ (MLE), the better the hypothesis is.

3. Dimension of parameter space: number of free parameters.

e.g. $H_0 : \theta = \theta_0 \in \Omega_0 \implies p = dim\Omega_0 = 0$, all parameters are fixed.

3. Thm: Under "regularity conditions",

$$-2\log\Lambda \xrightarrow{d} \chi^2_{p-d}$$

if H_0 is true, i.e. if $\theta \in \Omega_0$. ($p = dim\Omega$, $d = dim\Omega_0$)

4. Critical region: $R_\alpha = (\chi^2_{p-d,1-\alpha}, \infty)$
5. Unknown Variance: restricted likelihood (restrict to the null) V.S unrestricted likelihood (full parameter space).

⋆ $\sum_{i=1}^{n}(X_i - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \bar{X} + \bar{X} - \mu_0)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 + n(\bar{X} - \mu_0)^2$

6. Testing Independence:

⋆ In contingency table, if the row and column categories are independent, then $p_{ij} = P(Y_{ij} = 1) = p_{i.} \times p_{.j}$

9 Week 10: Power Sample Size Calculations

1. **Power** (η) = $P(Reject H_0 \mid H_0 false)$ = 1 - P(Type II Error), the probability of rejecting a false null hypothesis.

⋆ Tests with high power are able to detect deviations from H_0, and therefore "stronger".

⋆ Does not depend on the data.

2. Z-test Power:

$$\begin{aligned}
\eta &= P(\mid \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} \mid > z_{1-\alpha/2}) &(5)\\
&= 1 - P(-z_{1-\alpha/2} < \frac{\bar{X}-\mu_0}{\sigma_0/\sqrt{n}} < z_{1-\alpha/2}) &(6)\\
&= 1 - P(\frac{\mu_0-\mu_1}{\sigma_0/\sqrt{n}} - z_{1-\alpha/2} < \frac{\bar{X}-\mu_0+\mu_0-\mu_1}{\sigma_0/\sqrt{n}} < \frac{\mu_0-\mu_1}{\sigma_0/\sqrt{n}} + z_{1-\alpha/2}) &(7)\\
&= 1 - P(d\sqrt{n} - z_{1-\alpha/2} < Z < d\sqrt{n} + z_{1-\alpha/2}) &(8)\\
& &(9)
\end{aligned}$$

where **effect size** is $d = \frac{\mu_0-\mu_1}{\sigma_0}$

⋆ **Effect Size**: the number of standard deviations that μ_1 is away from μ_0 circumvents this problem while still retaining interpretability.

⋆ The power of Z-test to detect an effect of size d in a sample of size n, rejecting at the α significant level, is:
$\eta(d, n, \alpha) = 1 - (\Phi(d\sqrt{n} + z_{1-\alpha/2}) - \Phi(d\sqrt{n} - z_{1-\alpha/2}))$
(An interval of length $2z_{1-\alpha/2}$ under the normal curve, but shifted by $d\sqrt{n}$.)
d = 0 $\implies \eta(d, n, \alpha) = \alpha$
d or n $\to \infty \implies \eta(d, n, \alpha) = 1$
⋆ For same size n, the power to detect a larger effect d is larger than the power to detect a smaller effect.
⋆ For same effect d, the power to detect the effect is larger for large sample size n.

3. **Statistical** significance: what happened didn't just happen by luck, it might happen if we repeat the test. (used for rejecting the null.).
e.g. If we reject the null at 5% significant level, then we have observed a *statistically significant* deviation from μ_0 at this significant level.

Practically significant: you care what happened happened.
⋆ Statistical without practical: we saw sth completely meaningless, and we might see it again if we repeat the experiment.
⋆ Practical without statistical: we saw sth great but might not happen again if we repeat the experiment.
⋆ We need both to make a reasonable scientific conclusion.

4. Comparing Tests: the higher power the test has, the better the test is.

The rejection region of the t-test: $|t| = \mid \frac{\bar{X}-\mu_0}{s/\sqrt{n}} \mid > t_{1-\alpha/2}$

The rejection region of the likelihood ratio test: $-2\log\Lambda = n\log(1 + \frac{t^2}{n-1}) > \chi^2_{1-\alpha}$

⋆ LRT is a one sided test because of Λ instead of the χ^2: we reject the null if Λ is "small enough", i.e. $-2\log\Lambda$ is "large enough".

Thm (Neyman-Pearson Lemma): the likelihood ratio test is the most powerful. i.e. the test with critical region $\frac{f_1(x)}{f_0(x)} > c_\alpha$ has the higher power than any other tests. ????

⋆ **Uniformly most powerful**(composite alternatives): a test is uniformly most powerful if it is the most powerful against every possible simple alternatives.

⋆ The LRT is the UMP if there is one.

5. Sample Size Determination(Experiment design): we choose significant level, power and effect size to determine the sample size.
(1) Significant Level (α)

- common sense: don't make α too high such that any H_0 will be rejected
- depend on the field working
- empirical evaluation of the sensitivity of the procedure to this choice: evaluate the tradeoff between sample size, effect size and significant level and make sure that your experiment is robust to at least small changes in α

(2) The Power: be able to detect a deviation from H_0 with a certain probability.
(3) The Effect Size
(4) Use the power function $\eta(d, n, \alpha) = power$

$\star$ Tradeoff between Type I Error and Type II Error.

10 Week 11: Computational Methods: Jackknifes Bootstrap

1. **Standard Error**: the standard deviation of estimators. (usually these two terms are interchangeable.)

- Exact: when $X \sim N(\mu, \sigma^2)$ and $\hat{\mu} = \bar{X}$, then $SD(\hat{\mu}) = \sqrt{s/n}$
- Approximate: $SD(\hat{\theta}) \xrightarrow{CLT} \frac{1}{\sqrt{I(\theta_0)}}$
- if g is smooth, use Taylor series to linearize g.

$\star$ *Measures of Variability* are interpreted as the spread in values we would see in repeated sampling of a quantity.

2. Jackknife: (compute approximate standard error)
$\mathbf{x}_{(i)} = (x_1, ..., x_{i-1}), x_{i+1}, ..., x_n$, $\hat{\theta}_{(i)}$ is the estimator computed out of $\mathbf{x}_{(i)}$
Jackknife estimator: $\hat{SE}_{Jack} = (\frac{n-1}{n}\sum_{i=1}^{n}(\hat{\theta}_{(i)} - \hat{\theta}_{(.)})^2)^{1/2}$, where $\hat{\theta}_{(.)} = \frac{1}{n}\sum_{i=1}^{n}\hat{\theta}_{(i)}$

3. Non-Parameteric Bootstrap:

- Choose B $\in$ N
- for B in 1...B, obtain the bootstrap sample x_b by sampling n points from x, **with replacements**, then compute $\hat{\theta}_b = \hat{\theta}(x_b)$
- there might be duplicates due to "with replacement". $\rightarrow$ correlation.
- $\hat{F_\theta} \rightarrow \{x_b\}_{b=1}^{B} \rightarrow \{\hat{\theta}\}_{i=1}^{B}$

4. Parametric Bootstrap:

- Choose B $\in$ N
- for B in 1...B, obtain the bootstrap sample x_b by sampling n points from $F_{\hat{\theta}}$, then compute $\hat{\theta}_b = \hat{\theta}(x_b)$.
- $F_{\hat{\theta}} \rightarrow \{x_b\}_{b=1}^{B} \rightarrow \{\hat{\theta}\}_{i=1}^{B}$
- for hypithesis test: we have F_{θ_0}

5. Difference between Parametric and Non-Parameteric Bootstrap:
Non-parametric directly sample from the original sample, but parameteric assume the distribution of the sample first.