STAT 135: Concepts of Statistics

—Midterm Review Guide—

Disclaimer: This study guide was created by the Statistics Undergraduate Student Association, which is in no way affiliated with the current course staff of Statistics 135. Additionally, this guide is designed merely to *supplement* to your class notes.

Last Updated: Fall 2020

Originally Compiled by: Ethan Marzban

Contents

1 Inferential Statistics

- A **population** is too large to be observed in its entirety; as such, we must use **samples** to explore properties of the population.
  - The gold-standard of sampling is a **Simple Random Sample**, in which observations are taken independently of each other, from identical distributions.
  - Other sampling techniques include **stratified sampling**, and **cluster sampling.**
- **Population parameters** are deterministic quantities pertaining to the population; their exact values can never be determined exactly, and must be estimated using **estimators** (which are functions of data).
  - **Bias** measures "how far off" an estimator is from the parameter it is estimating. Mathematically, if $\hat{\theta}$ is an estimator for θ, we write $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}(\hat{\theta}) - \theta$.
- **Confidence Intervals** provide a way of quantifying uncertainty, specifically with respect to estimations of population parameters.
- The **Central Limit Theorem** provides information on the asymptotic behavior of the sample mean.
- Samples are generally taken *without* replacement, and observations are therefore (technically) dependent. However, as the sample size increases, the dependence between observations tends towards 0.
  - The **finite population correction factor** relates samples drawn without replacement to those drawn with replacement. When the sample size is very large, the finite population correction factor is approximately 1.

2 Parameter Estimation

- Two popular estimators of population parameters are the **Method of Moments (MoM)** estimator and the **Maximum Likelihood Estimator (MLE)**.
  - MLE's satisfy the **equivariance property** (sometimes called the **invariance property**), which states that $\widehat{f(\theta)}_{\text{ML}} = f(\hat{\theta}_{\text{ML}})$ (provided f satisfies certain mathematical properties).
  - The variance of the MoM estimator can be approximated using the **method of propagation of errors** (also known as the **δ-method**).
  - The asymptotic variance of the MLE is $[I_n(\theta)]^{-1}$, where $I_n(\theta)$ is the **Fisher information** of the sample.
- The **Cramér-Rao Lower Bound (CRLB)** provides a lower bound on the variance of *any* unbiased estimator of a population parameter (under certain regularity conditions).
  - Estimators whose variance are exactly equal to the CRLB are said to be **efficient**.
- **Sufficient Statistics** are functions (of data) that contain all the information about a parameter θ, given a sample. Mathematically, they induce a partition of the sample space that is finer than (or as fine as) the likelihood function. A **Minimal Sufficient Statistic** is a sufficient statistic that partitions the sample space in the *coarsest* manner.
  - Sufficient statistics are not unique, whereas minimally sufficient statistics are.[1]

[1] Technically, if T and U are both minimally sufficient statistics then there exists a one-to-one function ϕ such that $T = \phi(U)$, so it would be more mathematically rigorous to say $T \sim U$ rather than $T = U$.

- Sufficient statistics are typically found using the **Factorization Theorem**.

- An estimator $\hat{\theta}_n$ of θ is said to **converge in probability** to θ if $\mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) \to 0$ as $n \to \infty$. Convergence in probability is notated as: $\hat{\theta}_n \xrightarrow{\text{p}} \theta$.
  - Compare this with **convergence in distribution:** a sequence $\{X_i\}_{i=1}^n$ is said to **converge in distribution** to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

    Convergence in distribution is *weaker* than convergence in probability.

- An estimator $\hat{\theta}$ is said to be a **consistent** estimator of a parameter θ if $\hat{\theta} \xrightarrow{\text{p}} \theta$.
  - As an example: under certain regularity conditions, the MLE is consistent.

- Given an estimator $\hat{\theta}$ of a parameter θ, a "better" (i.e. lower-variance) estimator can always be obtained by conditioning on a sufficient statistic. This is known as **Rao-Blackwellization**.

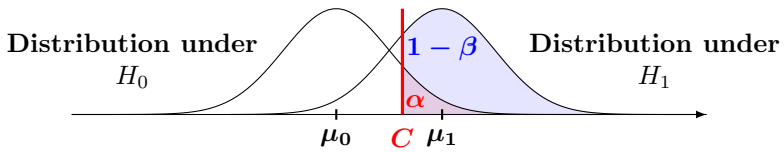
- The **bootstrap** provides another method for parameter estimation.
  - In the **nonparametric bootstrap**, no assumptions are made about the underlying distribution. The sampling distribution is approximated by repeatedly sampling (with replacement) from the original sample, and the remainder of inference is conducted as before.
  - In the **parametric bootstrap**, assumptions are made about the underlying distribution. The parameters of said distribution are computed from the original sample, and the sampling distribution is approximated by repeatedly generating samples from the assumed distribution (with the estimated parameters plugged in).

3 Hypothesis Testing

- The **null hypothesis** is chosen to represent the status quo; the **alternative hypothesis** provides a theory contrary to the null hypothesis. The goal of **hypothesis testing** is to determine which of the two hypotheses better describes the current state.

- There are several terms and notations associated with hypothesis testing:
  - **Level of Significance** (α): $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$
    - The event $\{(\text{reject } H_0 \mid H_0 \text{ is true}\}$ is known as a **Type I Error**
  - **Power [of a test]** ($1 - \beta$): $\mathbb{P}(\text{fail to reject } H_0 \mid H_0 \text{ is false})$
    - The event $\{(\text{fail to reject } H_0 \mid H_0 \text{ is false}\}$ is known as a **Type II Error**

Result of Test	State of H_0: **True**	**False**
Reject	**Type I**	No Error
Fail to Reject	No Error	**Type II**



- The **Neyman-Pearson Lemma** states that the **Generalized Likelihood Test** is **uniformly most powerful**.

- Hypothesis testing and confidence intervals are equivalent.

- $p-$values are always constructed "under the null;" that is, they are computed *after* assuming the null hypothesis is true.
- The **rejection region** is the set of values of a test statistic that lead to rejection of the null: $\mathcal{R} = \{x_i : \text{test at } x_i \text{ rejects } H_0\}$.

4 Partitioning the Sample Space

- A **partition** can be thought of a grouping structure that groups like elements together. For example, consider the likelihood function of our scenario above:

$$p(\theta) = \prod_{i=1}^{3} p(x_i;\theta) = \prod_{i=1}^{3} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{3} X_i}(1-\theta)^{3-\sum_{i=1}^{3} X_i}$$

- There are only **8** possible configurations of our data (X_1, X_2, X_3); let us enumerate them all:

$$\{(0,0,0),(0,0,1),(0,1,0),(1,0,0),(0,1,1),(1,0,1),(1,1,0),(1,1,1)\}$$

  The set of all possible configurations of data is called the **sample space**. So, the set above would be the sample space of our example.

- Each of these **8** possibilities corresponds to a different value of T:

(X_1, X_2, X_3)	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(1,0,0)$	$(0,1,1)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
$T(X_1, X_2, X_3)$	0	1	1	1	2	2	2	3

- Here, we can see how the statistic T induces a partition: it is natural to group all $T = 0$ terms together, all $T = 1$ terms together, and all $T = 2$ terms together.
- More generally, a partition is a collection of sets $\{B_1, \ldots, B_n\}$. A partition is defined to be **sufficient** if $f(x_i \mid X_1, \ldots, X_n \in B)$ does not depend on θ. By "the partition induced by T", we mean the partition consisting of elements $\{t : T(X_1, X_2, X_3) = t\}$ for all possible values of t. Therefore, for a statistic to be sufficient, the partition it induces must be sufficient.
- Algorithmically, to determine whether or not a *partition* is sufficient, we compare it to the partition induced by the conditional probability $\mathbb{P}(X_1, X_2, X_3 = (x_1, x_2, x_3) \mid T)$. For example,

$$p[(0,0,1) \mid T = 1] = \frac{\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0 \mid T = 1)}{\mathbb{P}(T = 1)} = \frac{\theta(1-\theta)^2}{\theta(1-\theta)^2 + \theta(1-\theta)^2 + \theta(1-\theta)^2} = \frac{1}{3}$$

  Iterating through the **8** possibilities, we obtain the following table:

(X_1, X_2, X_3)	$T(X_1, X_2, X_3)$	$p(X_1, X_2, X_3)$
(0, 0, 0)	0	$\frac{(1-\theta)^3}{(1-\theta)^3} = 1$
(0, 0, 1)	1	$1/3$
(0, 1, 0)	1	$1/3$
(1, 0, 0)	1	$1/3$
(0, 1, 1)	2	$1/3$
(1, 0, 1)	2	$1/3$
(1, 1, 0)	2	$1/3$
(1, 1, 1)	3	1

  Clearly the partition induced by $T = \sum X_i$ does not depend on θ; hence T is sufficient.

- As an example of a **non-sufficient** (or insufficient) statistic for θ, consider $T_2 = X_1$. Let us also examine one of the conditional probabilities in more detail:

$$p\left[(0,0,0) \mid T=0\right] = \frac{\mathbb{P}(X_1=0, X_2=0, X_3=0 \mid T=0)}{\mathbb{P}(T=0)} = (1-\theta)^2$$

Here is the full partition induced by this statistic:

(X_1, X_2, X_3)	$T_2(X_1, X_2, X_3)$	$p(X_1, X_2, X_3)$
(0, 0, 0)	0	$(1-\theta)^2$
(0, 0, 1)	0	$\theta(1-\theta)$
(0, 1, 0)	0	$\theta(1-\theta)$
(1, 0, 0)	1	$\theta(1-\theta)^2$
(0, 1, 1)	0	θ^2
(1, 0, 1)	1	$\theta(1-\theta)$
(1, 1, 0)	1	$\theta(1-\theta)$
(1, 1, 1)	1	θ^2

- Note that the **likelihood** itself induces a partition:

(X_1, X_2, X_3)	$L(\theta; X_1, X_2, X_3)$
(0, 0, 0)	$(1-\theta)^3$
(0, 0, 1)	$\theta(1-\theta)^2$
(0, 1, 0)	$\theta(1-\theta)^2$
(1, 0, 0)	$\theta(1-\theta)^2$
(0, 1, 1)	$\theta^2(1-\theta)$
(1, 0, 1)	$\theta^2(1-\theta)$
(1, 1, 0)	$\theta^2(1-\theta)$
(1, 1, 1)	θ^3

This so-called **likelihood partition** provides a good test for whether or not statistics are sufficient. If the likelihood partition has divisions (denoted by horizontal lines in the tables above) if is not sufficient. Otherwise, it is sufficient:

(X_1, X_2, X_3)	$T(X_1, X_2, X_3)$	$T_2(X_1, X_2, X_3)$	$L(\theta; X_1, X_2, X_3)$
(0, 0, 0)	0	0	$(1-\theta)^3$
(0, 0, 1)	1	0	$\theta(1-\theta)^2$
(0, 1, 0)	1	0	$\theta(1-\theta)^2$
(1, 0, 0)	1	1	$\theta(1-\theta)^2$
(0, 1, 1)	2	0	$\theta^2(1-\theta)$
(1, 0, 1)	2	1	$\theta^2(1-\theta)$
(1, 1, 0)	2	1	$\theta^2(1-\theta)$
(1, 1, 1)	3	1	θ^3

Note that the partition induced by T_2 "broke" a horizontal line; as such, it is not sufficient.

- A **minimal sufficient partition** is the coarsest sufficient partition. It can be shown that the likelihood always generates the coarsest sufficient partition; hence, we adopt the following "test:" a statistic is minimal sufficient (for θ) if the partition it induces is the same as the likelihood partition.

- In our example above, T_1 generates the same partition as the likelihood and is therefore **minimal sufficient**.

Rule-of-Thumb: If the likelihood partition creates divisions where the partition induced by T does not, then T is **not sufficient**; otherwise it is **sufficient**. If it T sufficient, and partitions the sample

space in the same way as the likelihood, then it is **minimal sufficient**.

5 A Useful Result

Theorem: Let $\hat{\theta}_n$ be an unbiased estimator of θ. If $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is a consistent estimator of θ.

Proof: Since $\hat{\theta}_n$ is an unbiased estimator of θ, we have that $\mathbb{E}(\theta_n) = \theta$. Let $\sigma_{\hat{\theta}_n}$ denote the standard deviation of $\hat{\theta}_n$ (for a fixed sample size n); then, Chebyshev's Inequality tells us that

$$0 \leq \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq k\sigma_{\hat{\theta}_n}\right) \leq \frac{1}{k^2} \tag{1}$$

for $k > 0$. Fix an arbitrary $\varepsilon > 0$; since equation 1 holds for any $k > 0$, it must hold for $k = \varepsilon/\hat{\theta}_n$, which yields

$$0 \leq \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \frac{\varepsilon}{\cancel{\sigma_{\hat{\theta}_n}}} \cdot \cancel{\sigma_{\hat{\theta}_n}}\right) \leq \frac{\sigma^2_{\hat{\theta}_n}}{\varepsilon^2}$$

Recognizing $\sigma^2_{\hat{\theta}_n} = \text{Var}(\hat{\theta}_n)$, and simplifying terms, we obtain

$$0 \leq \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \frac{\text{Var}(\hat{\theta}_n)}{\varepsilon^2}$$

Now, take the limit as $n \to \infty$; if $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then the squeeze theorem (from mathematics) tells us that

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \to 0$$

which is precisely the definition of the statement $\hat{\theta}_n \xrightarrow{\text{p}} \theta$. Therefore, we conclude that $\hat{\theta}_n$ is a consistent estimator of θ. □

6 Definitions and Theorem

Central Limit Theorem: Let $\{X_i\}_{i=1}^n$ denote a sample of size n, taken from a distribution with mean μ and standard deviation. Then

$$\bar{X} := \left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \to \mathcal{N}\left(\mu, \frac{\sigma^n}{n}\right) \qquad \text{as} \qquad n \to \infty$$

Theorem: Let $\hat{\theta}_{\text{ML}}$ denote the maximum likelihood estimator of a parameter θ. Then, as $n \to \infty$,

$$\hat{\theta}_{\text{ML}} \to \mathcal{N}\left(\theta, \frac{1}{nI(\theta)}\right)$$

where $I(\theta)$ denotes the Fisher Information:

$$I(\theta) = -\mathbb{E}\left[\frac{\partial^2}{\partial\theta^2}\ell(\theta)\right]; \qquad \ell(\theta) = \text{log-likelihood}$$

Wilks' Theorem: If Λ denotes the generalized likelihood ratio, then, as $n \to \infty$, $-2\log(\Lambda) \to \chi^2_{\text{df}}$ where $\text{df} = \dim(\Omega) - \dim(\omega_0)$.

Factorization Theorem: Consider a parameter θ and the associated likelihood function $L(x_1, \ldots, x_n; \theta)$. Then $T(x_1, \ldots, x_n)$ is a sufficient statistic for θ if and only if the likelihood factors as

$$L(x_1, \ldots, x_n; \theta) = g\left[T(x_1, \ldots, x_n), \theta\right] \cdot h(\theta)$$

Cramér-Rao Lower Bound: Given an i.i.d. sample $\{X_i\}_{i=1}^n$ with likelihood $f(x \mid \theta)$ and an unbiased estimator $T = t(x_1, \ldots, x_n)$, then (under certain regularity conditions)

$$\text{Var}(T) \geq \frac{1}{nI(\theta)}$$

Rao-Blackwell Theorem: Given an unbiased estimator $\hat{\theta}$ of θ with finite variance, and given statistic T that is sufficient for T, define $\tilde{\theta} := \mathbb{E}(\hat{\theta} \mid T)$. Then, for all θ,

$$\mathbb{E}[(\tilde{\theta} - \theta)^2] \leq \mathbb{E}[(\hat{\theta} - \theta)^2]$$

In other words, a lower-variance estimate can be obtained by conditioning an unbiased estimator on a sufficient statistic.

Efficiency: An estimator S of a parameter θ is said to be **efficient** if it attains the Cramér-Rao Lower Bound; that is, if

$$\text{Var}(S) = \frac{1}{nI(\theta)}$$

Consistency: An estimator $\hat{\theta}_n$ of a parameter θ is said to be **consistent** if $\hat{\theta}_n \xrightarrow{\text{p}} \theta$; that is, if

$$\lim_{n \to \infty} \left[\mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon)\right] = 0$$

for any $\varepsilon > 0$.

***p*−value:** The *p*−**value** of a test is defined to be the smallest level of significance at which a test would reject the null.

7 Exercises

Problem 1: Suppose I perform a series of $N = 100$ independent hypothesis tests, each at an α level of significance. Assuming that the null hypothesis is true, what is the chance that I reject at least one of these tests?

Hint: If R represents the number of tests I reject out of these 100, what distribution does R follow?

Problem 2: We have already seen that, given a sample $\{X_i\}_{i=1}^n$ from an $\text{Exp}(\lambda)$ distribution, $\hat{\lambda}_{\text{ML}} = 1/\bar{X}$. Let $\{Y_i\}_{i=1}^n$ denote a sample from an $\text{Exp}(1/\tau)$ distribution.

(a) With minimal computation, find $\hat{\tau}_{\text{ML}}$.

Hint: Use $\hat{\lambda}_{\text{ML}}$.

(b) With minimal computation, find $\hat{\tau}_{\text{ML}}$ and determine its **exact** distribution. That is, do **not** appeal to asymptotics.

Hint: What do we know about the sum of i.i.d. exponential distributions? How do exponential distributions scale?

Problem 3: The **Nakagami Distribution**, a distribution which bears many similarities with to Gamma distribution, has p.d.f. given by

$$f(x; m, \sigma) = \frac{2}{\Gamma(m)\sigma^m} x^{2m-1} \exp\left\{-\frac{x^2}{\sigma}\right\}; \qquad x \geq 0$$

(here, $m \geq 1/2$ and $\sigma > 0$ are population parameters). Let $\{X_i\}_{i=1}^n$ denote an i.i.d. sample from the Nakagami distribution with **known** parameter m and **unknown** parameter σ.

(a) Find $\hat{\sigma}_{\text{ML}}$, the maximum likelihood estimate of σ.

(b) The expressions for $\mathbb{E}(X)$ and $\text{Var}(X)$ are quite complicated; as such, let them be denoted simply by μ_X and s_X^2, respectively. Assuming the sample size n is very large, what distribution does $\hat{\sigma}_{\text{ML}}$ approximately follow?

Problem 4: Suppose $X \sim \text{Exp}(\lambda)$ where $\lambda > 0$ is an unknown population parameter.

(a) Find an expression for $\mathbb{E}(X^n)$ where $n \in \mathbb{R}^+$ is *not* necessarily an integer.

Hint: Start by writing out the integral, use Stat 134 tricks to avoid computing it directly.

(b) Let $v := \text{Var}(\sqrt{X})$. Find $\hat{v}_{\text{ML}}$, the maximum likelihood estimator of v.

Hint: Use equivariance.

Problem 5: Let $\{X_i\}_{i=1}^n$ denote an i.i.d. sample of size n from the $\text{Exp}(\lambda)$ distribution, where $\lambda > 0$ is an unknown parameter. Consider the following hypotheses:

$$\begin{cases} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda \neq \lambda_0 \end{cases}$$

where λ_0 is a known constant.

(a) Find a sufficient statistic for λ.

(b) Construct the GLRT (Generalized Likelihood Ratio Test).

(c) Sketch the power curve of the test you derived in part (a).

Hint: Trying to derive the mathematical expression for power will be difficult. Rather, use what you know about power to generate a rough "intuitive" sketch.

8 Answers

Problem 1: $R \sim \text{Binom}(100, \alpha) \implies \mathbb{P}(R \geq 1) = 1 - (1-\alpha)^{100}$

Problem 2: We have already seen that, given a sample $\{X_i\}_{i=1}^n$ from an $\text{Exp}(\lambda)$ distribution, $\hat{\lambda}_{\text{ML}} = 1/\bar{X}$. Let $\{Y_i\}_{i=1}^n$ denote a sample from an $\text{Exp}(1/\tau)$ distribution.

(a) $\hat{\tau}_{\text{ML}} = \bar{X}$.

(b) $\hat{\tau}_{\text{ML}} \sim \text{Gamma}\left(\tau, \frac{\tau^2}{n}\right)$.

Problem 3: **(a)** $\hat{\sigma}_{\text{ML}} = \frac{1}{nm}\sum_{i=1}^n x_i^2$

(b) $\hat{\sigma}_{\text{ML}} \sim \mathcal{N}\left(\sigma\ , \frac{1}{\frac{nm}{\sigma^2} - \frac{2n}{\sigma^3}\left(s_X^2+\mu_X^2\right)}\right) \overset{\text{d}}{=} \mathcal{N}\left(\sigma\ , \frac{\sigma^3}{n} \cdot \frac{1}{\sigma m - 2(s_X^2+\mu_X^2)}\right)$

Problem 4: Suppose $X \sim \text{Exp}(\lambda)$ where $\lambda > 0$ is an unknown population parameter.

(a) $\frac{\Gamma(n+1)}{\lambda^n}$

(b) $\hat{v}_{\text{ML}} = \frac{1}{\bar{X}} - \frac{1}{2}\sqrt{\frac{\pi}{\bar{X}}}$

Note: Recall that maximum likelihood estimators are *random variables*. As such, don't be frightened by the X in the final answer.

Problem 5: Let $\{X_i\}_{i=1}^n$ denote an i.i.d. sample of size n from the $\text{Exp}(\lambda)$ distribution, where $\lambda > 0$ is an unknown parameter. Consider the following hypotheses:

$$\begin{bmatrix} H_0 : \lambda = \lambda_0 \\ H_1 : \lambda \neq \lambda_0 \end{bmatrix}$$

where λ_0 is a known constant.

(a) $T(X) = \bar{X}e^{-\lambda\bar{X}}$.

(b) Rejection region $= \{\bar{X}e^{-\lambda_0\bar{X}} \leq c\}$

(c) Curve should be $u-$shaped, attaining a minimum value of α at λ_0. Curve should stop at the $y-$axis (rate parameters of exponential distributions cannot be negative), and tend asymptotically toward 1.