ESE 402/542 Midterm Review

Core Concepts So Far

- (Functions of) Random Variables
  - Mean, variance
  - Moments
- Central Limit Theorem
- Point Estimation
  - Fisher information, asymptotic variance, Cramer-Rao lower bound
  - Method of moments
  - Maximum likelihood
- Putting it all together: the beginnings of hypothesis testing
  - Type I, II errors – significance level, power
  - Rejection region, confidence intervals
  - (Generalized) likelihood ratio test

Goals of the Review Session

De-mystify the likely areas of confusion. Everything has a place in the puzzle!

Piece 1: The Central Limit Theorem

- The statement, plainly: the distribution of the sample mean looks more like a normal distribution as the number of samples increases.
- Main caveat (that we often take for granted): the random variables must have **bounded variance**.
- Mathematically: the distribution of $\bar{X}_n = \frac{1}{n} \sum_{i=1}^{n} X_i$ approaches $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$, where X_i are iid random variables with mean μ and variance σ^2.
- Notice that, equivalently, the distribution of $\sum_{i=1}^{n} X_i$ approaches a normal distribution $\mathcal{N}\left(n\mu, n\sigma^2\right)$.

Piece 1: The Central Limit Theorem

- Essentially, when you see a something related to a sum of iid random variables, the CLT should be on your mind.
- **The random variables do not have to be gaussian whatsoever**
- Simple example: What is the approximate distribution of $\sum_{i=1}^{n} X_i$, where $X_i \sim \text{Bernoulli}(p)$?
- Some of you might remember the rule of thumb that the Binomial distribution looks like Gaussian when n is large and p is not too close to 0 or 1. This confirms it.

Piece 2: CLT and the z-test

- Now that we've turned a complicated distribution of $\bar{X}_n$ into something that is approximately normal through CLT, this allows us to provide a general framework for doing statistics on $\bar{X}_n$.
- Recall that if X_i have mean μ and variance σ^2, then $\bar{X}_n \overset{\text{approx}}{\sim} \mathcal{N}(\mu, \sigma^2/n)$.
- z-test statistic is $z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$

Piece 2: CLT and the z-test

- Quite a few people have asked about about the normalizing factor on the denominator.
  - There's no mathematical difference to consider $\bar{X}_n - \mu$ instead. At the end of the day what we want to quantify is the *number of standard deviations away* $\bar{X}_n$ is from the true mean μ.
  - $z = \frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}}$ ensures that $z \sim \mathcal{N}(0, 1)$–this allows us to use a single table of numbers in the back of a textbook rather than an infinite number.

Piece 2: CLT and the *z*-test

- The *z*-test as stated assumes we know the true variance σ^2.
- When we don't know σ^2, what is the natural thing to do? Estimate it, of course! This is a recurring theme in statistics.
- When we use the estimated variance $\hat{\sigma}^2$, technically we are using the *t*-test, but for *n* large enough, the *t*-distribution is essentially identical to the normal distribution, hence why we don't mention it in this class.

Piece 3: Beyond the Mean – Parameters in general

- What is a parameter of a random variable? Intuitively, something that tells you something about the distribution of a random variable. Sometimes, only tells part of the story e.g. μ for normal distribution $\mathcal{N}(\mu, \sigma^2)$, other times fully characterizes a distribution e.g. p for Bernoulli(p).
- Examples of parameters are mean, variance, higher-order moments etc.
- Parameters can be redundant. E.g. $\mathbb{E}[X] = \mu$, $\mathbb{E}\left[X^2\right] := c$, $\text{var}(X) = \mathbb{E}\left[X^2\right] - \mathbb{E}[X]^2 = c - \mu^2$.

Piece 3: Beyond the Mean – Parameters in general

- Since we have a lot of choices for parameters θ, what determines which one we use for statistical testing?
  - We certainly want one that whose distribution we can understand approximately (e.g. through CLT). This is why we use the sample mean so often. This allows us to evaluate confidence intervals and probabilities (e.g. $\bar{X}_n$ is 2 standard deviations away from $\mu \implies < 5\%$ probability of happening by chance).
  - Converse question to ask: how informative is our data in determining our parameter? Intuitively, if our data's information is low for the parameter, then we'll need a *lot* of samples to get a good estimate of the parameter. This is underlying idea of Fisher information, and subsequently Cramer-Rao lower bound. If your data's information is low, then the variance of your parameter estimate is going to be high!

Piece 4: Tools for Estimating Parameters

- Method of Moments. Intuitively, the moments of a distribution often contain a lot of information about distribution (e.g. gaussian distribution).
  - Use the fact that $\sum_{i=1}^{n} X_i^k$ is (unbiased) estimator of k-th moment $\mathbb{E}\left[X^k\right]$ to write parameter of interest (e.g. variance) as function of moments, which we can estimate from data.
  - This is an algebraic approach! No part of this procedure worries about information lower bound, unbiasedness of resulting estimator etc.
- Maximum likelihood. Intuitively, given the data that you've observed, what is the parameter that maximizes the *likelihood* of your data occurring. Makes sense, right?
  - We have shown that the asymptotic variance of MLE is Cramer-Rao lower bound–this also makes sense: we are using our "best" estimate of the parameter from the data.
  - However, definitely not unbiased in most cases: intuitively, we are overfitting to the observed data.

Piece 5: Hypothesis Testing

- In abstraction, we have a null hypothesis H_0 and alternative hypothesis H_1.
- Want to gather statistical evidence to reject null hypothesis – this is different from "proving H_1 true" or "proving H_0 false".
- Set a "significance level" α. Intuitively, an acceptable risk parameter. If an outcome has less than α probability of occurring via H_0 by *chance*, then we say that's enough evidence to reject H_0 in favor of H_1.

Piece 5: Hypothesis Testing

- This leads to concept of Type I and Type II errors. Type I error α is by design the significance level: "what is probability that I reject my null hypothesis H_0 when my data actually comes from H_0?". Type II error β: "what is probability I fail to reject my null hypothesis when my data doesn't come from H_0?"
- p-value. Related to type I error. Let's say $\alpha = 0.05$ for H_0, where $X \sim \mathcal{N}(0, 1)$. Two-sided rejection region is $|z| \geq 1.96$. p-value is the probability of statistic of your data occurring via H_0: let's say $\bar{X} = 3$, then p-value is ≈ 0.003
- Notion of power: $1 - \beta$. In other words, what is probability I reject my null hypothesis when I'm supposed to?
- Can't re-iterate importance of power enough! Only caring about p-values can lead to malicious statistics!