MATH 321: Test 3 Study Guide

Lecture 5 – The Central Limit Theorem (5.6 and 5.7)

Convergence in distribution

- Definition: A sequence of random variables, $Y_1, Y_2, \ldots$, converges in distribution to a random variable Y if $\lim_{n\to\infty} F_{Y_n}(y) = F_Y(y)$ at all points y where $F_Y(y)$ is continuous (notation: $Y_n \xrightarrow{d} Y$).

CLT

Central Limit Theorem: Let $X_i \overset{iid}{\sim} f(x)$ with $E(X) = \mu$ and $V(X) = \sigma^2 > 0$. Then the distribution of

$$W = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \text{Normal}\,(0,1) \quad \text{as } n \to \infty$$

- Normal mgf theorem: If $Z \sim \text{N}\,(\mu = 0, \sigma^2 = 1)$, and μ and $\sigma > 0$ are constants, then $X = \sigma Z + \mu \sim \text{N}\,(\mu, \sigma^2)$
- Results of CLT
  (a) $\frac{\sigma}{\sqrt{n}} W + \mu = \bar{X}$ can be approximated by $\frac{\sigma}{\sqrt{n}} Z + \mu \sim \text{Normal}\,(\mu, \frac{\sigma^2}{n})$ for "large" n.
  (b) $n\bar{X} = X_1 + \ldots + X_n = S$ can be approximated by $(\sigma\sqrt{n})Z + n\mu \sim \text{Normal}\,(n\mu, n\sigma^2)$ for "large" n.

t, Z, and the CLT

- If $X_1, \ldots, X_n$ are a random sample for a $N(\mu, \sigma^2)$, as $n \to \infty$, $t_{n-1} \xrightarrow{d} Z$
- If $X_1, \ldots, X_n$ are not normal random variables, when the sample size is large

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \overset{approx}{\sim} \text{Normal}\,(0,1) = Z \quad \text{by CLT}$$

Normal approximation to discrete distributions

- Continuity correction: If $X \sim$ Discrete with corresponding $S \sim$ Normal by the CLT, then for integers $a \le b$:

$$P(X = a) = P(a - 0.5 \le S \le a + 0.5) \qquad \text{and} \qquad P(a \le X \le b) = P(a - 0.5 \le S \le b + 0.5)$$

- Normal approximation to binomial
  - Result: If $X \sim \text{Binomial}\,(n, p) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = np, \sigma^2 = npq)$
  - Conditions: $np \ge 5$ and $nq = n(1-p) \ge 5$
- Normal approximation to Poisson
  - Result: If $X \sim \text{Poisson}\,(\lambda) \Longrightarrow X \approx S \sim \text{Normal}\,(\mu = \lambda, \sigma^2 = \lambda)$
  - Condition: $\lambda \ge 10$

Central interval probabilities

- Empirical rule: If $X \overset{approx}{\sim} \text{Normal}$, then
  1. Approximately 68% of data is within $\mu \pm \sigma$
  2. Approximately 95% of data is within $\mu \pm 2\sigma$
  3. Approximately 99.7% of data is within $\mu \pm 3\sigma$

Lecture 6 – Confidence Intervals (7.1 - 7.4)

Interval estimators / confidence intervals

- Definition: An interval estimator or confidence interval for how to calculate endpoints of an interval from sample data: $[L(\mathbf{X}), U(\mathbf{X})]$

  Once $\mathbf{X} = \mathbf{x}$ is observed, the inference $L(\mathbf{x}) \leq \theta \leq U(\mathbf{x})$ is made.
- Goals of CIs: (1) Capture the target parameter θ (2) Be relatively narrow
- Confidence coefficients definition:

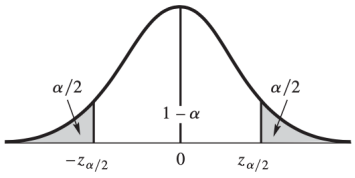
  Probability that a CI captures $\theta \rightarrow P\big(L(\mathbf{X}) \leq \theta \leq U(\mathbf{X})\big) = 1 - \alpha$ for significance level α
- $100(1-\alpha)\%$ CI for $\theta = [L(\mathbf{X}), U(\mathbf{X})]$

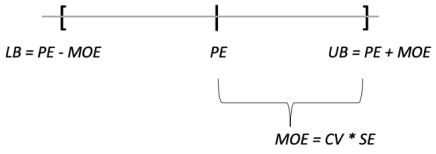
Constructing confidence intervals

- Setup: $\hat{\theta}$ = unbiased point estimator for parameter θ;
  $\sigma_{\hat{\theta}}$ = standard deviation of the sampling distribution of $\hat{\theta}$ (i.e. standard error of $\hat{\theta}$)

  If $\hat{\theta} \sim \text{Normal}\,(\theta, \sigma_{\hat{\theta}})$ (or approximately normal) $\implies Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \sim \text{Normal}\,(0,1)$
- To find interval for θ with confidence coefficient equal to $1-\alpha$, need critical values $-z_{\alpha/2}$ and $z_{\alpha/2}$ such that $P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) = 1-\alpha$. Then

$$
\begin{aligned}
1-\alpha &= P(-z_{\alpha/2} \leq Z \leq z_{\alpha/2}) \\
&= P(-z_{\alpha/2} \leq \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \leq z_{\alpha/2}) \\
&= P(\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}} \leq \theta \leq \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}) \\
&\implies 100(1-\alpha)\%\ \text{CI} \ = [\hat{\theta} - z_{\alpha/2}\,\sigma_{\hat{\theta}}, \hat{\theta} + z_{\alpha/2}\,\sigma_{\hat{\theta}}] \\
&= \hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}
\end{aligned}
$$

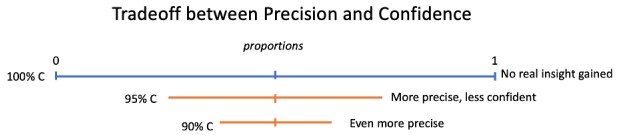




- Summary of CIs
  - Point Estimate (PE) is the best guess; at the center of the interval.
  - Margin of Error (MOE) = Critical Value (CV) × Standard Error (SE).

    SE (standard deviation of the statistic) measures sampling error.

    % Confident is determined by confidence level set and incorporated via the critical value (CV).
- All else equal, here is how the researcher can affect the precision of intervals:
  - Larger sample size $n \rightarrow$ smaller interval (smaller SE)
  - More confident $\rightarrow$ larger interval (larger CV)



- Interpretation general structure:

  I am % confident that the true/population parameter + context is between lower bound and upper bound.

Types of intervals

- Variables that affect the form of intervals:
  - Independent or dependent samples
  - Sample sizes n_1 and n_2 (large or small)
  - Population distributions X_1 and X_2 (normal or not normal)
  - Population variances σ_1^2 and σ_2^2 (known or unknown and ratio of variances)
- Large sample confidence intervals

  If n is large $\implies \hat{\theta} \overset{approx}{\sim} \text{Normal}\,(\theta, \sigma_{\hat{\theta}}) \implies 100(1-\alpha)\%$ CI $= \hat{\theta} \pm z_{\alpha/2}\,\sigma_{\hat{\theta}}$

  Conditions: for means $n_i \geq 30$; for proportions $n_i p_i \geq 5$ and $n_i(1-p_i) \geq 5$

θ	$\hat{\theta}$	$\sigma_{\hat{\theta}}$	
μ	$\bar{X}$	$\dfrac{\sigma}{\sqrt{n}}$	Estimate σ^2 with s^2 if unknown
$\mu_1 - \mu_2$	$\bar{X}_1 - \bar{X}_2$	$\sqrt{\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}}$	Estimate σ_i^2 with s_i^2 if unknown
p	$\hat{p}$	$\sqrt{\dfrac{p(1-p)}{n}}$	Estimate p with $\hat{p}$
$p_1 - p_2$	$\hat{p}_1 - \hat{p}_2$	$\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}$	Estimate p_i with $\hat{p}_i$

  If starting from $X_i \sim$ Normal and known variances, then intervals are exact; if not then $X_i \approx$ Normal by CLT and confidence coefficients are approximate.

- Small sample confidence intervals for means ($n_i < 30$)

  If n is small $\implies 100(1-\alpha)\%$ CI $= \hat{\theta} \pm t_{\alpha/2}\,\sigma_{\hat{\theta}}$ $\qquad$ t crit values $>$ z crit values

  Conditions: for one sample $X \sim$ Normal with unknown σ^2; for two samples $X_1 \perp\!\!\!\perp X_2$ and $X_1, X_2 \sim$ Normal with unknown common variance σ^2

Parameter	*Confidence Interval ($\nu = df$)*
μ	$\overline{Y} \pm t_{\alpha/2}\left(\frac{S}{\sqrt{n}}\right), \qquad \nu = n-1.$
$\mu_1 - \mu_2$	$(\overline{Y}_1 - \overline{Y}_2) \pm t_{\alpha/2} S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}},$ where $\nu = n_1 + n_2 - 2$ and $S_p^2 = \frac{(n_1-1)S_1^2 + (n_2-1)S_2^2}{n_1+n_2-2}$ (requires that the samples are independent and the assumption that $\sigma_1^2 = \sigma_2^2$).

  If not starting from $X_i \sim$ Normal then confidence coefficients are approximate and work well as long as not badly skewed with outliers.

- Dependent samples CI for $\mu_1 - \mu_2$

  Simplifies to a one sample CI for the differences $\mu_1 - \mu_2 = \mu_D$ shown above if n is small

- One-sided CI

Lower bound (at least)	Upper bound (at most)
$P(\hat{\theta} - z_\alpha\,\sigma_{\hat{\theta}}) = 1-\alpha$	$P(\hat{\theta} + z_\alpha\,\sigma_{\hat{\theta}}) = 1-\alpha$
$\implies [\hat{\theta} - z_\alpha\,\sigma_{\hat{\theta}}, \infty)$	$\implies (-\infty, \hat{\theta} + z_\alpha\,\sigma_{\hat{\theta}}]$

Margin of error (MOE) revisited

- $MOE = \frac{UB-LB}{2} = \frac{Width}{2} \qquad \rightarrow \qquad Width = 2 \times MOE$
- The **error in estimation ϵ** is the distance between an estimator and its target parameter:
  $[\hat{\theta} - \epsilon, \hat{\theta} + \epsilon] \implies |\hat{\theta} - \theta| = \epsilon$

Finding minimum sample size

- We want the $100(1-\alpha)\%$ confidence interval for θ, $\hat{\theta} \pm z_{\alpha/2}\sigma_{\hat{\theta}}$, to be no longer than that given by $\hat{\theta} \pm \epsilon$, then for
  - One mean μ with $V(X) = \sigma^2$ known and $X \sim$ Normal or assume going to have "large" n:

    $n \geq \frac{z_{\alpha/2}^2 \sigma^2}{\epsilon^2}$

    If σ^2 is unknown, use best approximation available.
  - One proportion p: $n \geq \frac{z_{\alpha/2}^2\, p^*(1-p^*)}{\epsilon^2}$

    If there is prior knowledge, use $p^* = \hat{p}$, else set $p^* = 0.5$

Lecture 7 – Hypothesis Tests (8.1 - 8.3)

Hypothesis test

- Definition: A hypothesis test is a rule that specifies

  For which sample value the decision is made to reject H_0 in favor of H_A.

  For which sample value the decision is made to "not reject" H_A in favor of H_A.

- Elements of a hypothesis test

  1. Null hypothesis H_0 and Alternative hypothesis H_A

     Definitions:

     – Hypotheses are statements about population parameters

     – The Null hypothesis H_0 is an assumption about θ that is assumed to be true

     – The Alternative hypothesis H_A is the complement of H_0

  2. Test statistic (TS) and Rejection Region RR

     TS: Function of the sample $W(X_1, \ldots, X_n)$, think of this as the point estimator $\hat{\theta}$

     RR: Subset of the sample space (range of sample) for which H_0 will be rejected

  3. Conclusion / interpretation

     General structure: Because our test statistic (COMPARISON of TS and RR) (IS / IS NOT) in the rejection region we (REJECT or FAIL TO REJECT) the null hypothesis.
     At the (ALPHA) significance level, there (IS or IS NOT) sufficient evidence to conclude (THE ALTERNATIVE HYPOTHESIS).

Large sample tests

- If n is large, then $\hat{\theta} \sim \text{Normal}\,(\theta, \sigma_{\hat{\theta}})$ (or approximately normal) $\Longrightarrow Z = \frac{\hat{\theta}-\theta}{\sigma_{\hat{\theta}}} \sim \text{Normal}\,(0,1)$

- Using the same parameters θ, point estimates $\hat{\theta}$, and standard errors $\sigma_{\hat{\theta}}$ as shown in confidence intervals, all of the large sample α-level tests can be summarized with

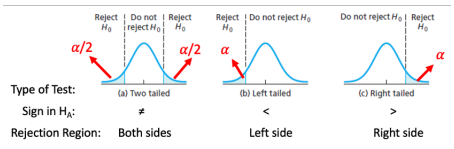
Large-Sample α-Level Hypothesis Tests

$H_0 : \theta = \theta_0$.

$H_a : \begin{cases} \theta > \theta_0 & \text{(upper-tail alternative).} \\ \theta < \theta_0 & \text{(lower-tail alternative).} \\ \theta \neq \theta_0 & \text{(two-tailed alternative).} \end{cases}$

Test statistic: $Z = \dfrac{\hat{\theta} - \theta_0}{\sigma_{\hat{\theta}}}$.

Rejection region: $\begin{cases} \{z > z_\alpha\} & \text{(upper-tail RR).} \\ \{z < -z_\alpha\} & \text{(lower-tail RR).} \\ \{|z| > z_{\alpha/2}\} & \text{(two-tailed RR).} \end{cases}$



- For proportions

  – One sample: In the standard error, use $p_0 \Longrightarrow \sigma_{\hat{p}} = \sqrt{\dfrac{p_0(1-p_0)}{n}}$.

  – Two sample: In the standard error, use $p_1 = p_2 = p$ and estimate with $\hat{p} = \dfrac{x_1 + x_2}{n_1 + n_2} \Longrightarrow \sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\hat{p}(1-\hat{p})[1/n_1 + 1/n_2]}$.

- In any particular test, only one of the listed alternatives H_A is appropriate, which will be based on the research question. Then use the corresponding rejection region.

Small sample tests for means

- If n is small, then need to switch to t-tests. For these we start with $X \sim$ Normal
- Summary of the small-sample α-level tests for μ

A Small-Sample Test for μ

Assumptions: $Y_1, Y_2, \ldots, Y_n$ constitute a random sample from a normal distribution with $E(Y_i) = \mu$.

$H_0 : \mu = \mu_0$.

$$H_a: \begin{cases} \mu > \mu_0 & \text{(upper-tail alternative).} \\ \mu < \mu_0 & \text{(lower-tail alternative).} \\ \mu \neq \mu_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $T = \dfrac{\overline{Y} - \mu_0}{S/\sqrt{n}}$.

$$\text{Rejection region: } \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases} \quad t_\alpha, \text{ with df} = n - 1$$

- If we are testing two independent means $\mu_1 - \mu_2$ and assume both Normal distributions with common unknown variance σ^2, then we use the pooled variance S_p^2 as the estimator for σ^2 in the standard error $\sigma_{\bar{X}_1 - \bar{X}_2}$. Then

Small-Sample Tests for Comparing Two Population Means

Assumptions: Independent samples from normal distributions with $\sigma_1^2 = \sigma_2^2$.

$H_0 : \mu_1 - \mu_2 = D_0$.

$$H_a: \begin{cases} \mu_1 - \mu_2 > D_0 & \text{(upper-tail alternative).} \\ \mu_1 - \mu_2 < D_0 & \text{(lower-tail alternative).} \\ \mu_1 - \mu_2 \neq D_0 & \text{(two-tailed alternative).} \end{cases}$$

Test statistic: $T = \dfrac{\overline{Y}_1 - \overline{Y}_2 - D_0}{S_p\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}$, where $S_p = \sqrt{\dfrac{(n_1 - 1)S_1^2 + (n_2 - 1)S_2^2}{n_1 + n_2 - 2}}$.

$$\text{Rejection region: } \begin{cases} t > t_\alpha & \text{(upper-tail RR).} \\ t < -t_\alpha & \text{(lower-tail RR).} \\ |t| > t_{\alpha/2} & \text{(two-tailed RR).} \end{cases}$$

Here, $P(T > t_\alpha) = \alpha$ and degrees of freedom $\nu = n_1 + n_2 - 2$.

- If we are testing dependent means $\mu_1 - \mu_2$, then have paired t-test

  $\mu_1 - \mu_2 = \mu_D \quad \rightarrow \quad \frac{\bar{D} - \mu_D}{S_D/\sqrt{n}} = T \sim \mathrm{t}_{n-1}$, one sample test on differences

  If not starting from $X_i \sim$ Normal then t-tests are approximately α-level and work well as long as not badly skewed with outliers.

p-values

- Definition: A p-value is the probability that under the null hypothesis the test statistic will be at least as "extreme" as the observed value.
- Two ways to make conclusion for hypothesis tests:

  Traditional method: $TS \overset{?}{\in} RR$

  p-value method: Reject H_0 if p-value $\leq \alpha$ and Fail to reject H_0 if p-value $> \alpha$

Relationship between confidence intervals and hypothesis tests

- Confidence interval = Acceptance region = Complement of RR
- Decisions based on CI: For $H_0 : \theta = \theta_0$ and

  Two-tailed $H_A : \theta \neq \theta_0$: "Accept" H_0 if θ_0 falls within the $100(1-\alpha)\%$ CI and reject if outside.

  Right-tailed $H_A : \theta > \theta_0$: Reject if outside lower bound CI

  Right-tailed $H_A : \theta < \theta_0$: Reject if outside upper bound CI

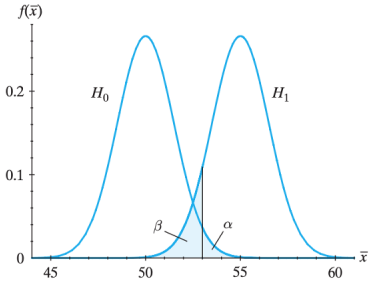
Type I and Type II errors

- Type I: Incorrectly rejecting H_0

  $\alpha = P(\text{Type I error}) = P(\text{Reject when } H_0 \text{ is true}) = P(TS \in RR \mid H_0)$
- Type II: Incorrectly failing to reject H_0

  $\beta = P(\text{Type II error}) = P(\text{Fail to reject when } H_0 \text{ is false}) = P(TS \notin RR \mid H_A)$

Example for one mean



- Power: Correctly rejecting H_0

  Power $= 1 - \beta = P(\text{Reject } H_0 \text{ when } H_0 \text{ is false}) = P(TS \in RR \mid H_A)$