Lecture 2

3. Inference for other percentiles
   a. Suppose \( \theta \) is quantile \( r \in (0, 1) \) of distribution of iid \( X_1, \ldots, X_n \).
      i. That is, \( P \{ X_j \leq \theta \} = r \).
   b. Generalization of Sign Test:
      i. Let \( T \sim \text{Bin}(n, \tau) \).
      ii. Let \( T \) be number of observations smaller than \( \theta \).
      iii. Under \( H_0 \), \( T \sim \text{Bin}(n, \tau) \).
   c. Power is \( \sum_{j=0}^{b-1} \tau^j(1-\tau)^{n-j}(\theta) \geq 1 - \alpha \).
   d. Often largest \( a \) so that \( \sum_{j=0}^{a-1} \tau^j(1-\tau)^{n-j}(\theta) < \alpha/2 \).
      \( a \) is \( \alpha/2 \) quantile of \( \text{Bin}(n, \tau) \) distribution
   e. Power is \( \sum_{j=0}^{b-1} \tau^j(1-\tau)^{n-j}(\theta) < \alpha/2 \).
   f. Power is \( \sum_{j=0}^{a-1} \tau^j(1-\tau)^{n-j}(\theta) < \alpha/2 \).
   g. Reject \( H_0 \) if \( T < a \) or \( T > b \).
   h. Confidence Interval for \( \theta \) is \( (X_{(a)}, X_{(b)}) \).

E. Comparing Tests
   1. For fixed size, alternative, and power, one with smaller sample size is better.
      a. Make numeric comparison by taking ratio of these two sample sizes: Relative efficiency.
   b. Suppose that we have
      i. Parameter \( \theta \) with null value \( \theta^0 \).
      ii. Alternative is \( \theta > \theta^0 \).
      iii. Two families (dependent on sample size) of one-sided test statistics \( T_1 \) and \( T_2 \).
      iv. Test size \( \alpha \), power \( 1 - \beta \).
   c. Compare implied sample sizes
      i. Pick \( n_1 \) sample size for test 1.
      ii. Pick \( \theta \) so that test \( T_1 \) of size \( \alpha \) has power \( 1 - \beta \) for alternative \( \theta \).
      iii. Pick \( n_2 \) so that test \( T_2 \) of size \( \alpha \) has power \( 1 - \beta \) for alternative \( \theta \).
   d. Easy case: Asymptotically normal statistics are compared using standard deviations and derivatives of means under alternative.
      a. \( T_1 \sim N(\mu_1(\theta), \sigma_1^2(\theta)/n_1) \), approximately.
      b. Critical value: Find \( c \) such that \( P_0 \{ T_1 \geq c \} = \alpha \)
        \( \mu_1(\theta) = \mu_1(0) + \mu_1'(0)\theta, \sigma_1(\theta) = \sigma_1(0) \).

Lecture 2

f. This interval isn’t of the form that you are used to:
   \( \hat{\theta} \pm 2\hat{\sigma} \), for \( \hat{\sigma} \) with a factor of \( 1/\sqrt{n} \).
   i. \( \text{Var}[\text{median}] \approx 1/4(\frac{1}{\text{var}})^2 \) (Cramér, H. (1946) Mathematical Methods of Statistics, Princeton U. Press p. 368f. Cramér proves asymptotic normality using an argument that I’m not sure is complete.)
   ii. We will learn later on how to estimate this, but it’s harder than the earlier CI rule.

11 Lecture 2

d. Note that confidence level is conservative:
   i. \( P \{ X_{(a)} \leq \theta \leq X_{(b)} \} = 1 - P \{ X_{(a)} \geq \theta \} \geq 1 - \alpha \).
   ii. For any given \( \theta \), inequality is generally strict.

F. Estimating the CDF of \( X_1, \ldots, X_n \)
   1. identically distributed
2. For \( x \) in the range of \( X_j \), let \( \tilde{F}(x) \) be the number of data points less than or equal to \( x \), divided by \( n \).

3. If observations are independent, \( F(x) \sim n^{-1} \text{Bin}(n, F(x)) \).
   a. Confidence interval for \( F(x) \) is
\[
\tilde{F}(x) \pm z_{a/2} \sqrt{\frac{\tilde{F}(x)(1-\tilde{F}(x))}{n}} \text{ or exact version.}
\]
   b. CLT intervals will extend outside \([0,1]\), which is not reasonable.
      i. Use exact, Truncate, or take 542 to learn other approaches.

### IV. Two-Sample Testing

#### A. Independent Samples
1. Data \( Y_1, \ldots, Y_n \) from continuous CDF \( G \)
2. Data \( X_1, \ldots, X_m \) from continuous CDF \( F \)
3. Null hypothesis: distributions are identical: \( F(y) = G(y) \forall y \).
4. Usual Approach: Two-sample pooled \( t \) test
   a. Pooled variance estimate
   
   \[ T = \frac{(n+m-2)s_p^2}{s_p^2} \]
   
   \[ s_p^2 = \frac{\sum_{j=1}^m (X_j - \bar{X})^2 + \sum_{j=1}^n (Y_j - \bar{Y})^2}{(m+n-2)} \]

5. Nonparametric approach analogous to sign test: Mood’s Median Test.
   a. Calculate combined sample median
   b. Let \( A \) be number of observations from \( Y \) ’s above the combined median

\[
\begin{array}{ccc}
  Y & X & \text{Total} \\
  \text{> Median} & B & (m+n-1)/2 \\
  = \text{Median} & C & 1 \\
  \text{< Median} & A & (m+n-1)/2 \\
  \text{Total} & n & m & m+n \\
\end{array}
\]

Use as test statistic \( B - A \).
   i. \( P[B - A \geq t] = P[B - A \geq t | C = 0] \frac{m+n}{(m+n)} P[B-A \geq t | C = 1] \frac{n}{(m+n)} \)
   h. Called Mood’s Median Test.

6. Nonparametric approach: Rank sum statistic
   a. Assume data are continuous.
   b. Rank all of the observations
      i. Under null hypothesis, all orderings have equal probabilities
   c. Compute \( W = \) sum of ranks in one of the groups (M, perhaps)
   d. Calculate expectation and variance
      i. \( W = \sum_{j=1}^{m+n} I_j j \) for
         \[ I_j = \begin{cases} 1 & \text{if subject ranked } j \text{ is from } G \\ 0 & \text{otherwise.} \end{cases} \]
         \[ E[I_j] = n/(m+n) \]
      ii. \( E[W] = \frac{n}{m+n} \sum_{j=1}^{m+n} j = \frac{n(m+n)(m+n+1)}{2(m+n)} \)
      iii. Variance is harder, since the \( I_j \) are not independent.
   e. If \( V \) is the sum of ranks for the other group, then
   \[ V + W = (n+m)(n+m+1)/2 \]
   f. Compare against normal.

### Lecture 2

#### 6. Nonparametric approach: Rank sum statistic
   a. Assume data are continuous.
   b. Rank all of the observations
      i. Under null hypothesis, all orderings have equal probabilities
   c. Compute \( W = \) sum of ranks in one of the groups (M, perhaps)
   d. Calculate expectation and variance
      i. \( W = \sum_{j=1}^{m+n} I_j j \) for
   
   \[ I_j = \begin{cases} 1 & \text{if subject ranked } j \text{ is from } G \\ 0 & \text{otherwise.} \end{cases} \]
      ii. \( E[I_j] = n/(m+n) \)
      iii. \( E[W] = \frac{n}{m+n} \sum_{j=1}^{m+n} j = \frac{n(m+n)(m+n+1)}{2(m+n)} \)
      iv. Variance is harder, since the \( I_j \) are not independent.
   e. If \( V \) is the sum of ranks for the other group, then
   \[ V + W = (n+m)(n+m+1)/2 \]
   f. Compare against normal.

### Lecture 3

#### 6. Nonparametric approach: Rank sum statistic
   a. Assume data are continuous.
   b. Rank all of the observations
      i. Under null hypothesis, all orderings have equal probabilities
   c. Compute \( W = \) sum of ranks in one of the groups (M, perhaps)
   d. Calculate expectation and variance
      i. \( W = \sum_{j=1}^{m+n} I_j j \) for
   
   \[ I_j = \begin{cases} 1 & \text{if subject ranked } j \text{ is from } G \\ 0 & \text{otherwise.} \end{cases} \]
      ii. \( E[I_j] = n/(m+n) \)
      iii. \( E[W] = \frac{n}{m+n} \sum_{j=1}^{m+n} j = \frac{n(m+n)(m+n+1)}{2(m+n)} \)
      iv. Variance is harder, since the \( I_j \) are not independent.
   e. If \( V \) is the sum of ranks for the other group, then
   \[ V + W = (n+m)(n+m+1)/2 \]
   f. Compare against normal.