Mathematical Statistics Recitation 10

TA: Sonia Reilly

April 10, 2026

Lecture Review

Lecture 17: Two Sample Testing, Normally Distributed

Normal data, σ unknown

$$X_1, \dots, X_n \sim N(\mu_X, \sigma^2), \quad Y_1, \dots, Y_m \sim N(\mu_Y, \sigma^2)$$
$$H_0 : \mu_X = \mu_Y, \qquad H_1 : \mu_X \neq \mu_Y$$

Two-sample t-test rejects for large T: (denominator is an estimate of $\mathrm{Var}(\bar{X}_n - \bar{Y}_m)$)

$$\sqrt{-2\log\Lambda} \approx T = \frac{\bar{X}_n - \bar{Y}_m}{\left[\left(\frac{1}{n} + \frac{1}{m}\right)\frac{1}{n+m-2}\left(\sum(X_i - \bar{X}_n)^2 + \sum(Y_j - \bar{Y}_m)^2\right)\right]^{1/2}} \sim t_{n+m-2} \quad \text{under } H_0$$

Higher power $1-\beta$ if: larger $|\mu_X - \mu_Y|$, larger α, larger m, n, lower variance σ^2

Normal data, σ known, $m = n$

$$\mathrm{Var}(\bar{X}_n - \bar{Y}_n) = \frac{2\sigma^2}{n}$$

$$\frac{\sqrt{n}(\bar{X}_n - \bar{Y}_n)}{\sqrt{2}\sigma} \sim N(0,1) \quad \Longrightarrow \quad \text{reject if} \quad |\bar{X}_n - \bar{Y}_n| \geq \sqrt{\frac{2}{n}}\sigma z_{\alpha/2}$$

Can compute power exactly.

Non-normal data For large m, n, $T \to N(0,1)$ (still assumes same variance σ^2), but power unknown.

Lecture 18: Two Sample Testing, Nonparametric

Independent data: Mann–Whitney Test Sort $(X_1, \dots, X_n, Y_1, \dots, Y_m)$ lowest to highest, rank 1 to $m+n$.

$$T_Y = \sum_j \mathrm{rank}(Y_j)$$

For small m, n, can compute distribution exactly. For large m, n:

$$\frac{T_Y - \mathbb{E}[T_Y]}{\sqrt{\mathrm{Var}(T_Y)}} = \frac{T_Y - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} \xrightarrow{d} N(0,1) \quad \text{under } H_0$$

Reject if absolute value $\geq z_{\alpha/2}$.

Paired Samples (X_i, Y_i) independent, X_i not independent from Y_i (e.g., same test subject, before/after treatment).

If data is normal: t-test

$$D_i = X_i - Y_i, \quad \bar{D}_n = \frac{1}{n}\sum D_i$$

$$\frac{\sqrt{n}\,\bar{D}}{\sqrt{\frac{1}{n-1}\sum(D_i - \bar{D}_n)^2}} \sim t_{n-1}$$

Having paired samples increases efficiency (lowers $\mathrm{Var}(\bar{X}_n - \bar{Y}_n)$) if X_i's, Y_i's positively correlated

If data is not normal (nonparametric): Signed Rank Test

$$D_i = X_i - Y_i$$

Order $|D_i|$ from lowest to highest.

$$W_+ = \sum_{D_i > 0} \mathrm{rank}(|D_i|)$$

$$\frac{W_+ - \mathbb{E}[W_+]}{\sqrt{\mathrm{Var}(W_+)}} = \frac{W_+ - \frac{n(n+1)}{4}}{\sqrt{\frac{n(n+1)(2n+1)}{24}}} \xrightarrow{d} N(0,1) \quad \text{under } H_0$$

Reject if absolute value $\geq z_{\alpha/2}$.

Lecture 19: Testing Differences in Populations for Categorical Data

Fisher's Exact Test

$$\begin{array}{cc|c} N_{11} & N_{12} & n_{1\cdot} \\ N_{21} & N_{22} & n_{2\cdot} \\ \hline n_{\cdot 1} & n_{\cdot 2} & n_{\cdot\cdot} \end{array}$$

$$H_0: \ P(\text{row } 1 \mid \text{col } 1) = P(\text{row } 1 \mid \text{col } 2), \quad P(\text{row } 2 \mid \text{col } 1) = P(\text{row } 2 \mid \text{col } 2)$$

Treat column and row sums as fixed, so N_{11} determines the whole distribution (hypergeometric):

$$P(N_{11}) = \frac{\binom{n_{1\cdot}}{N_{11}}\binom{n_{2\cdot}}{N_{21}}}{\binom{n_{\cdot\cdot}}{n_{\cdot 1}}}$$

Can calculate pmf exactly and compute 0.05 level rejection region.
Exact, don't need $n \to \infty$. Impractical for large # samples or ≥ 2 categories. When to use: if $n_{\cdot\cdot}$ is very small, but mostly only if it's explicitly asked for in the problem

Chi-Squared Test of Homogeneity J distributions/columns, I categories/rows
Column sums fixed. Q: Are the distributions the same?

$$H_0: \pi_{i1} = \cdots = \pi_{iJ} \quad \forall i, \qquad H_A: \text{not } H_0$$

Shown in class that generalized LR test gives:

$$\hat{\pi}_i = P(\text{cell } i) = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \qquad \Longrightarrow \qquad E_{ij} = \frac{n_{\cdot j} n_{i\cdot}}{n_{\cdot\cdot}}$$

$$T = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(n_{ij} - n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot})^2}{n_{i\cdot}n_{\cdot j}/n_{\cdot\cdot}} \longrightarrow \chi^2_{(I-1)(J-1)}$$

Chi-Squared Test of Independence Q: is the column variable independent of the row variable? Column assignment is random, unlike above.

$$H_0 : \pi_{ij} = \pi_{i\cdot}\pi_{\cdot j} \quad \forall i, j, \qquad H_A : \pi_{ij} \text{ are free}$$

Same test statistic as before, but for a different question and setup. Comes from different MLE:

$$\hat{\pi}_{ij} = \hat{\pi}_{i\cdot}\hat{\pi}_{\cdot j} = \frac{n_{i\cdot}}{n_{\cdot\cdot}} \cdot \frac{n_{\cdot j}}{n_{\cdot\cdot}} \qquad \Longrightarrow \qquad E_{ij} = \frac{n_{\cdot j} n_{i\cdot}}{n_{\cdot\cdot}}$$

Problems

1. [11.21] A study was done to compare the performances of engine bearings made of different compounds. Ten bearings of each type were tested. The following table gives the times until failure (in units of millions of cycles):

ranks ↘	X Type I	Y Type II	
1	3.03	3.19	2
8	5.53	4.26	3
9	5.60	4.47	4
11	9.30	4.53	5
13	9.92	4.67	6
14	12.51	4.69	7
17	12.95	12.78	16
18	15.21	6.79	10
19	16.04	9.37	12
20	16.84	12.75	15

To avoid boring arithmetic, you may use the following facts (where Type I are the X_i's and Type II are the Y_i's):

$$\bar{X} = 10.69 \qquad \bar{Y} = 6.75$$

$$\sum(X_i - \bar{X})^2 = 209.03 \qquad \sum(Y_i - \bar{Y})^2 = 116.80.$$

(a) Test the hypothesis that there is no difference between the two types of bearings, under the assumption that both distributions are normal.

(b) Test the same hypothesis using a non-parametric method.

(c) Which of the methods – that of part (a) or that of part (b) – do you think is better in this case? (Hint: Think about our assumption in part (a). Does it seem correct?)

Solution:

(a) Assuming the samples are normally distributed, we use the two-sample t-test with $m = n$:

$$T = \frac{\bar{X} - \bar{Y}}{\sqrt{\frac{1}{n(n-1)}\left[\sum(X_i - \bar{X})^2 + \sum(Y_j - \bar{Y})^2\right]}} \sim t_{2n-2}$$

$$T = 2.07$$

$$\text{p-value} = P(|T| \geq 2.07) \quad \text{for } T \sim t_{18} \;\Rightarrow\; \text{p-value} = 0.053$$

(b) Use Mann–Whitney. Find the ranks by ordering the data.

$$T = \sum_j \text{rank}(Y_j) = 2 + 3 + 4 + 5 + 6 + 7 + 16 + 10 + 12 + 15 = 80$$

$$\frac{T - \mathbb{E}[T]}{\sqrt{\text{Var}(T)}} = \frac{T - \frac{m(m+n+1)}{2}}{\sqrt{\frac{mn(m+n+1)}{12}}} = \frac{80 - 105}{\sqrt{\frac{100(21)}{12}}} = -1.89$$

$$\text{p-value} = P(|Z| \geq 1.89) = 0.064$$

(c) Mann–Whitney is better — the data does not seem to fit the assumption that it comes from a normal distribution very well (particularly looking at Type II, there's a few big outliers.)

2. The presence of an environmental factor is thought to increase the incidence of a disease. Based on the data, does it make a difference? Use a test of significance level 0.05.

	not present	present
no disease	128	7
disease	19	7

Solution:

Chi-squared test (of independence, since both variables are random):

$$E_{ij} = \frac{n_{\cdot j} n_{i \cdot}}{n_{\cdot\cdot}}$$

Compute:

$$n_{\cdot 1} = 147, \quad n_{\cdot 2} = 14, \quad n_{1 \cdot} = 135, \quad n_{2 \cdot} = 26, \quad n_{\cdot\cdot} = 161$$

$$E = \begin{pmatrix} 123.26 & 11.74 \\ 23.74 & 2.26 \end{pmatrix}$$

$$T = \sum_{i=1}^{I} \sum_{j=1}^{J} \frac{(O_{ij} - E_{ij})^2}{E_{ij}} = \frac{(128 - 123.26)^2}{123.26} + \frac{(7 - 11.74)^2}{11.74} + \frac{(19 - 23.74)^2}{23.74} + \frac{(7 - 2.26)^2}{2.26} = 12.98$$

$$(I - 1)(J - 1) = (2 - 1)(2 - 1) = 1$$

So $T \to \chi^2_1$.

From table, p-value is < 0.005, so we reject the null hypothesis at significance level 0.05, i.e., we decide that the environmental factor does make a difference (technically, the data would be very unlikely if the disease and the factor were independent).

3. [13.29] Suppose a company wishes to examine the relationship of gender to job satisfaction, grouping job satisfaction into four categories: very satisfied, somewhat satisfied, somewhat dissatisfied, and very dissatisfied. The company plans to ask the opinions of 100 employees. Should you, the company's statistician, carry out a chi-square test of independence or a test of homogeneity?

   Solution:

   It depends on how the sampling is done.

   (a) If a predetermined number of each gender is surveyed, use homogeneity.

   (b) If 100 employees are chosen at random from the whole employee body, use independence.

4. [13.28] In a chi-squared test of homogeneity, would the results change if instead of counts, the entries were proportions of the column total, with each column summing to 1? What if the entries were percentages of the column total?

   Solution:

   The results would change! We would no longer know how many individuals were sampled, so intuitively we should not be able to determine if the results are statistically significant without this info.

   More mathematically, this means replace n_{ij} with $n_{ij}/n_{\cdot j}$, so

$$\text{MLE } \hat{\pi}_i = \frac{\sum_j (n_{ij}/n_{\cdot j})}{\sum_{i,j}(n_{ij}/n_{\cdot j})}$$

   The new O_{ij} is

$$O_{ij} = \frac{n_{ij}}{n_{\cdot j}}$$

   and the new E_{ij} is

$$E_{ij} = \left(\frac{\sum_j (n_{ij}/n_{\cdot j})}{\sum_{i,j}(n_{ij}/n_{\cdot j})} \right) n_{\cdot j}$$

   The $n_{\cdot j}$ does not cancel and the answer is different.

   Using percentages would give yet a third answer — E_{ij} and O_{ij} would each be 100 times larger than in the proportion case, so

$$\frac{(O_{ij} - E_{ij})^2}{E_{ij}}$$

   would be 100 times larger, and the test statistic value would be 100 times larger, thus the p-value would be smaller (i.e., the test thinks there's 100 times as many data points).

5. [11.19] An experiment is planned to compare the mean μ_X of a control group to the mean μ_Y of an independent sample of a group given a treatment. Suppose that the observations are approximately normally distributed and that the standard deviation of a single measurement in either group is $\sigma = 5$. There are $n = 25$ samples in each group.

   (a) What is the standard error of the difference of the sample means, $\overline{Y} - \overline{X}$?

(b) With a significance level $\alpha = 0.05$, what is the rejection region of the test of the null hypothesis $H_0 : \mu_Y = \mu_X$ versus the alternative $H_A : \mu_Y > \mu_X$?

(c) What is the rejection region of the level $\alpha = 0.05$ test if the alternative is $H_A : \mu_Y \neq \mu_X$?

Solution:

(a)

$$S_{\bar{Y}-\bar{X}} = \sqrt{\text{Var}(\bar{Y} - \bar{X})} = \sqrt{\frac{2\sigma^2}{n}} = \sqrt{2}$$

(b)

$$\bar{Y} - \bar{X} \geq \sqrt{2} z_{0.05} = \sqrt{2}(1.65) = 2.33$$

(c)

$$|\bar{Y} - \bar{X}| \geq \sqrt{2} z_{0.025} = \sqrt{2}(1.96) = 2.78$$

6. [11.24] Find the exact null distribution of the Mann–Whitney test statistic T for $m = 3$, $n = 2$.

Solution:

Under H_0, each permutation of $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ is equally likely to be the correct increasing order. There are

$$\binom{m+n}{m} = \binom{5}{3} = 10$$

possibilities for the ranks of the Y_j's. Below are the permutations and their corresponding statistic

$$T = \sum_j \text{rank}(Y_j) :$$

$$\begin{array}{lll} 1+2+3=6 & 1+3+5=9 & 2+4+5=11 \\ 1+2+4=7 & 1+4+5=10 & 3+4+5=12 \\ 1+2+5=8 & 2+3+4=9 & \\ 1+3+4=8 & 2+3+5=10 & \end{array}$$

So, the pmf of T is

t	6	7	8	9	10	11	12
$P(T=t)$	$\frac{1}{10}$	$\frac{1}{10}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{5}$	$\frac{1}{10}$	$\frac{1}{10}$