Lecture 23
Outline

Nonparametric Statistics
  Parametric vs Nonparametric
  Robustness
  Wilcoxon’s Signed Rank Test
  Wilcoxon Rank Sum Test
  Spearman’s Correlation

Brief Review for Final Exam
Part I

Nonparametric Statistics
In analyzing random data, we rarely if ever know their “true” distribution, and the “true” distribution would not likely be simple anyway.

For ignorance and/or convenience, we often assume the data are normally distributed, and therefore can be characterized by two quantities: the mean $\mu$ and variance $\sigma^2$.

The inferential methods we have discussed have almost all revolved around means like $\mu$ (except for the “crummy $F$-test”).
Central Limit Theorem

- Fortunately, even when the data are not normally distributed, the Central Limit Theorem (CLT) guarantees that the sample mean $\bar{X}$ has a normal sampling distribution provided the sample size is large enough (typically $n \geq 30$).

- Most of the inferential methods we have seen (all the way through MLR) somehow make use of this normal distribution (and its unknown-$\sigma^2$ variant, the $t$-distribution).

- But what if the normal assumption for $\bar{X}$ is not valid – say, when the underlying population is not normal and the sample size $n$ is small?
Outliers

- Biggest reasons to doubt the normality of our data: outliers.
- When present, the data collection process should be re-examined.
- Outliers can only be completely discarded when they are known to be errors.
- Otherwise, do separate analyses with and without the outliers and compare the results.
- If they differ substantially (influential points), think about the science; the outliers could be the most informational data points! (Example: Challenger O-ring data: the two failures at very low temperatures were outliers in the dataset, yet provided most of the [tragically ignored] relevant information.)
Strategies for Non-normality

- Do nothing. *t*-distribution-based procedures are rather *robust*, and thus relatively safe even under mild violations of normality.

- Do nothing but say something. Discuss possible consequence of non-normality when it is of some concern.

- Use a transformation to make the data look more normal (though this sometimes makes interpretation of the results more difficult).

- If known, use the actual distribution of the data, e.g., the exact test for binomial data (see Lecture 14).

- *Use a nonparametric procedure*....
Nonparametrics

.stopPropagation()“Classical” statistics are based theoretical distributions with a few (finite number of) parameters. Inference based on these distributions are known as parametric.

Nonparametric inference does not assume a specific mathematical form (like a normal or a $t$ or whatever) for the distribution of the data.

So “nonparametric” really means “infinitely parametric.”

Nonparametric inference is more robust, but can also be less powerful.

Nonparametric procedures are usually more computationally difficult, and perhaps prohibitively so, even with modern computing.
Robustness

- To measure the uncertainty in our inferences, we use $p$-values and $100 \times (1 - \alpha)\%$ confidence intervals.
- Both of these associate a probability with some quantities computed from the data.
- The probability ($p$ or $1 - \alpha$) is estimated and hence is referred to as the nominal level.
- A robust procedure should estimate that probability accurately (e.g., get actual CI coverage very close to the nominal level, $1 - \alpha$) over a wide array of data scenarios (including, say, a normal distribution contaminated with outliers).
Power

- One test is more powerful than another if it yields a smaller \textit{p-value} (more significance) for the same data, when the same alternative is true.
- One CI procedure is more powerful than another if it yields \textit{narrower intervals} for the same nominal coverage level $1 - \alpha$.
- There is often a \textit{trade-off} between robustness and power.
Rank-based Tests

Three rank-based (i.e., replacing the data values by their ranks) nonparametric procedures described in the remainder of this lecture:

▶ Wilcoxon Signed Rank Test: for one-sample inference; analog of one-sample t test, M&M Section 7.1.
▶ Wilcoxon Rank Sum Test: for two-sample inference; analog of two-sample t test, M&M Section 7.2.
▶ Spearman (nonparametric) correlation coefficient: for measuring bivariate association; analog of r, the usual Pearson (parametric) correlation coefficient, M&M Section 2.2.

NOTE: The book’s version of this material is actually on the CD-ROM that should have been attached to the inside back cover!
One-Sample Inference

As in Lecture 11, consider the case of paired samples, $X_i$ and $Y_i$:

- Recall we often set $D_i = X_i - Y_i$, so that $\mu_D = \mu_X - \mu_Y$, and desire a test of

$$H_0 : \mu_X = \mu_Y \iff \mu_D = 0$$

- Another way of expressing the null hypothesis is that “$D$ is equally likely to be $-a$ or $a$;” a negative deviation of amount $a$ is as likely as a positive deviation of amount $a$.

- That is, we assume the $D_i$ are symmetric about 0, but make no other assumptions about their distribution.
Wilcoxon Signed Rank Test

- Order the absolute differences $|D_1|$, $|D_2|$, $\ldots$, $|D_n|$ from smallest to largest. If any of the $D_i = 0$, discard them.
- Let $R_i$ denote the rank of $|D_i|$. If there are ties among the $|D_i|$, assign each the average of the ranks they occupy.
- Let $W^+ = \sum R_i$ for all $i$ such that $D_i > 0$. This is the Wilcoxon Signed Rank statistic.
- Under $H_0$, $W^+$ has mean and variance
  \[
  \mu_{W^+} = \frac{n(n + 1)}{4} \quad \text{and} \quad \sigma_{W^+}^2 = \frac{n(n + 1)(2n + 1)}{24}.
  \]
- Note that $\mu_{W^+}$ equals half of the sum of all the ranks.
- Reject the null when $W^+$ is too far from its mean.
Example

Hamilton Depression Scale Factor IV data
These data, based on nine patients who received a certain tranquilizer, were taken from a double-blind clinical trial involving two tranquilizers. The measure used was the Hamilton (t960) depression scale factor IV (the “suicidal” factor) at the first or second visit.
# Hamilton Depression Data

<table>
<thead>
<tr>
<th>Patient ($i$)</th>
<th>Pre ($Y_i$)</th>
<th>Post ($X_i$)</th>
<th>diff ($D_i$)</th>
<th>sign</th>
<th>rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>1.83</td>
<td>0.88</td>
<td>−0.95</td>
<td>−1</td>
<td>8</td>
</tr>
<tr>
<td>2</td>
<td>0.50</td>
<td>0.67</td>
<td>0.17</td>
<td>1</td>
<td>3</td>
</tr>
<tr>
<td>3</td>
<td>1.62</td>
<td>0.60</td>
<td>−1.02</td>
<td>−1</td>
<td>9</td>
</tr>
<tr>
<td>4</td>
<td>2.48</td>
<td>2.05</td>
<td>−0.43</td>
<td>−1</td>
<td>4</td>
</tr>
<tr>
<td>5</td>
<td>1.68</td>
<td>1.06</td>
<td>−0.62</td>
<td>−1</td>
<td>7</td>
</tr>
<tr>
<td>6</td>
<td>1.88</td>
<td>1.29</td>
<td>−0.59</td>
<td>−1</td>
<td>6</td>
</tr>
<tr>
<td>7</td>
<td>1.55</td>
<td>1.06</td>
<td>−0.49</td>
<td>−1</td>
<td>5</td>
</tr>
<tr>
<td>8</td>
<td>3.06</td>
<td>3.14</td>
<td>0.08</td>
<td>1</td>
<td>2</td>
</tr>
<tr>
<td>9</td>
<td>1.30</td>
<td>1.29</td>
<td>−0.01</td>
<td>−1</td>
<td>1</td>
</tr>
</tbody>
</table>
Wilcoxon Signed Rank results for the Depression data

- \( W^+ = 5 \), \( \mu_{W^+} = 22.5 \), and \( \sigma_{W^+} = \sqrt{71.25} = 8.44 \). Note that \( W^+ \) is more than 2 standard deviations below its null mean.
- For small samples, SAS can compute the \( p \)-value of this test **exactly** through extensive table look-ups; for large samples, a normal approximation is OK!
- In this example, we get \( Z = (5 – 22.5)/8.44 = -2.07 \). So from Table A, this implies a 2-sided \( p \)-value of
  \[
  p = 2(.0192) = .0384.
  \]
- SAS gets the **exact** answer, \( p = 0.0391 \).
```markdown
Nonparametric Statistics

Parametric vs Nonparametric
Robustness
Wilcoxon's Signed Rank Test
Wilcoxon Rank Sum Test
Spearman's Correlation

### data depression;
```input x y @@;
D = x - y;
datalines;
```
+-----------+-----------+-------+--------+
<table>
<thead>
<tr>
<th>D</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.83 0.878 0.50  0.674</td>
</tr>
<tr>
<td>1.62 0.598 2.48  2.05</td>
</tr>
<tr>
<td>1.68 1.06  1.88  1.29</td>
</tr>
<tr>
<td>1.55 1.06  3.06  3.14</td>
</tr>
<tr>
<td>1.30 1.29</td>
</tr>
</tbody>
</table>
```
```
### Tests for Location: \( \mu_0 = 0 \)

<table>
<thead>
<tr>
<th>Test</th>
<th>-Statistic-</th>
<th>-----p Value------</th>
</tr>
</thead>
<tbody>
<tr>
<td>Student’s t</td>
<td>t 2.981831</td>
<td>( \text{Pr} &gt; \</td>
</tr>
<tr>
<td>Sign</td>
<td>M 2.5</td>
<td>( \text{Pr} \geq \</td>
</tr>
<tr>
<td>Signed Rank</td>
<td>S 17.5</td>
<td>( \text{Pr} \geq \</td>
</tr>
</tbody>
</table>
Two-Sample Inference

- Now suppose two groups are independently randomized to two drugs to compare the outcomes $X$ and $Y$.

- Chapter 7 hypotheses: $H_0 : \mu_X = \mu_Y$ vs. $H_a : \mu_X = \mu_Y + \Delta$.

- More general, nonparametric formulation: $H_0 : F(X) = G(Y)$ where $F$ and $G$ are unspecified distributions, vs. $H_a : F(X) = F(Y + \Delta)$.

- That is, $H_a$ corresponds only to a shift in the location of the distribution, not a change in its shape. Still, we make no further assumptions about $F$ at all.
Wilcoxon Rank Sum Test

- Pool the two independent datasets $X_1, \ldots, X_{n_1}, Y_1, \ldots, Y_{n_2}$, and rank all $N = n_1 + n_2$ observations together.
- Let $R_i$ denote the resulting ranks; if there are ties, again use average ranks.
- Let $W = \sum R_i$ for all $i$ in the $X$ sample. This is the Wilcoxon Rank Sum statistic (also sometimes called the Mann-Whitney statistic).
- Under $H_0$, $W$ has mean and variance
  \[ \mu_W = \frac{n_1(N + 1)}{2} \quad \text{and} \quad \sigma_W^2 = \frac{n_1 n_2 (N + 1)}{12}. \]
- Reject $H_0$ when $W$ is too far from its null mean.
**Example**

Glucose levels in healthy and lead-poisoned geese!

<table>
<thead>
<tr>
<th>Healthy Geese (X)</th>
<th>Lead-Poisoned Geese (Y)</th>
</tr>
</thead>
<tbody>
<tr>
<td>value</td>
<td>combined rank</td>
</tr>
<tr>
<td>297</td>
<td>8</td>
</tr>
<tr>
<td>340</td>
<td>12</td>
</tr>
<tr>
<td>325</td>
<td>10</td>
</tr>
<tr>
<td>227</td>
<td>1</td>
</tr>
<tr>
<td>277</td>
<td>3</td>
</tr>
<tr>
<td>337</td>
<td>11</td>
</tr>
<tr>
<td>250</td>
<td>2</td>
</tr>
<tr>
<td>290</td>
<td>5</td>
</tr>
</tbody>
</table>

The SAS Proc here is called **NPAR1WAY**
Wilcoxon Rank Sum results for the Geese data

\[ W = 8 + 12 + 10 + 1 + 3 + 11 + 2 + 5 = 52, \]
\[ \mu_W = 8(15 + 1)/2 = 64, \text{ and} \]
\[ \sigma^2_W = (8)(7)(15 + 1)/12 = 74.667. \]

As for the one-sample Wilcoxon test, for small samples SAS can compute an exact \( p \)-value for this test; for large samples, a normal approximation is again used.

In this example, we get \( Z = (52 - 64)/\sqrt{74.667} = -1.39. \) So from Table A, this implies a 2-sided \( p \)-value of

\[ p = 2(.0823) = .1646. \]

So not significant at \( \alpha = 0.05 \) this time.

After swapping the roles of X and Y (which doesn’t matter), SAS gets the exact answer, \( p = 0.1893. \)
data goose;
  input glucose group @@;
datalines;
297 1 340 1 325 1 227 1
277 1 337 1 250 1 290 1
293 2 291 2 289 2 430 2
510 2 353 2 318 2 ;
proc npar1way data = goose;
  class group;
  var glucose;
  exact;
run;
Wilcoxon Two-Sample Test

Statistic (S) 68.0000

Normal Approximation
Z 1.3309
One-Sided Pr > Z 0.0916
Two-Sided Pr > |Z| 0.1832

t Approximation
One-Sided Pr > Z 0.1023
Two-Sided Pr > |Z| 0.2045

Exact Test
One-Sided Pr >= S 0.0946
Two-Sided Pr >= |S - Mean| 0.1893
Spearman’s Correlation

- Pearson’s correlation $r$ measures the **linear** association between two variables.
- $r$-based inference assumes **normal** distributions for $X$ and $Y$.
- If we order the two sets of data *separately*, Spearman’s rank correlation is simply the Pearson correlation between the two sets of **ranks**.
- Spearman’s correlation $r_{Spearman}$ measures the **monotonic** association between two variables.
- **Example**: If $y_i = x_i^2$ for $x_i = 0, 1, 2, 4, 6$, then $r \approx 0.7$ but $r_{Spearman} = 1.0!$
- Spearman’s correlation can be used when one or both of the variables are merely **ordinal**, instead of quantitative.
SAS code for Pearson, Spearman correlations

```sas
proc corr data = depression spearman pearson;
   var x y;
run;
```
# SAS output, Pearson

Pearson Correlation Coefficients, N = 9
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1.00000</td>
<td>0.84342</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0043</td>
</tr>
<tr>
<td>y</td>
<td>0.84342</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0043</td>
</tr>
</tbody>
</table>
SAS output, Spearman

Spearman Correlation Coefficients, N = 9
Prob > |r| under H0: Rho=0

<table>
<thead>
<tr>
<th></th>
<th>x</th>
<th>y</th>
</tr>
</thead>
<tbody>
<tr>
<td>x</td>
<td>1.00000</td>
<td>0.64708</td>
</tr>
<tr>
<td></td>
<td></td>
<td>0.0596</td>
</tr>
<tr>
<td>y</td>
<td>0.64708</td>
<td>1.00000</td>
</tr>
<tr>
<td></td>
<td>0.0596</td>
<td></td>
</tr>
</tbody>
</table>
Ties

- All rank-based procedures get more complicated when there are ties among the ranks, which typically reduce the information content of the data.

- Software will use the appropriate (and often approximate) ranking and \( p \)-value-computing procedures when there are ties.
Part II

Brief Review for Final Exam
Lectures 15-16

- **Odds Ratio**: definition, contrast with relative risk, use in case-control studies
- **Chi-squared test of association**: observed vs. expected counts under the null hypothesis of independence, equivalence to two-sample binomial test in $2 \times 2$ table case, subsequent generalization to $r \times c$ case
- **Paired $2 \times 2$ tables**: McNemar’s test, redefinition of odds ratio
- **Measuring agreement**: $\kappa$ statistic
- **Diagnostic tests**: sensitivity, specificity, positive predictive value, and negative predictive value
Lectures 17-18

▶ *Simpson’s Paradox*: Marginal association can be different from each conditional association, perils of data aggregation

▶ *Covariance and Correlation*: definitions, properties, estimation, interpretation; impact of nonlinearity, nonnormality, and outliers

▶ *Simple Linear Regression (SLR) basics*: definition, four key assumptions (independence, normality, homoscedasticity, linearity), interpretation, computation via least squares
Lectures 19-20

- **Connection between Correlation and SLR**: Deriving SLR slope $\hat{b}$ and intercept $\hat{a}$ from $r$, $s_x$, $s_y$, $\bar{x}$, and $\bar{y}$; “variance explained” interpretation of $r^2$

- **SLR Inference**: annoying notational change (to $\beta$’s) in Ch 11, estimation, CI’s and hypothesis tests via the $t_{n-2}$ distribution, fitted values $\hat{Y}$, $\text{MSE} = \hat{\sigma}^2$

- **SLR Prediction**: of mean response $\mu_y$ at some $X = x$, or of a new observation $Y$; $\hat{y} = \hat{\mu}_y$ but $(SE_{\hat{y}})^2 = (SE_{\hat{\mu}})^2 + \text{MSE}$; confidence bands versus prediction bands

- **SLR Diagnostics**: scatterplots of the residuals and what assumptions they can test, normal q-q plots of the residuals, influential points and Cook’s distance
Lectures 21-22

- **Multiple Linear Regression (MLR) Basics**: multiple predictors, estimation, inference, interpretation, collinearity
- **ANOVA Table for MLR**: relationships among elements in the table and to the multiple $R^2$
- **Model Fit and Evaluation**: $R^2$, diagnostic plots
- **Variable Selection**: via stepwise elimination of insignificant predictors, or maximization of $R^2$ over all models
- **Special Topics**: transformations (especially log and quadratic; comparisons among), binary predictors, dummy variables (treatment and accumulative contrasts versus numerical scoring methods), and interactions (potential advantage over additive models, interpretation vis-a-vis main effects)
Lecture 23

- **Nonparametrics Basics**: Comparison with parametric approach, impact of outliers on normality, robustness versus power
- **One-sample Nonparametric Inference**: Wilcoxon’s Signed Rank Test
- **Two-sample Nonparametric Inference**: Wilcoxon’s Rank Sum Test
- **Spearman’s Correlation**: Pearson correlation $r$ applied to the ranks of the data