Dispersion or Variability

How Much Do Distributions or Data Vary?

Slides used by Prof. Charles Peters
More Variable – Less Variable

Numeric
More Variable – Less Variable
Categorical
Variability of Numeric Variables

- Variance: \( \sigma^2 = E[(X-\mu)^2] \),
  \[ s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \overline{x})^2 \]
- Standard deviation: \( \sigma = \sqrt{\sigma^2} \), \( s = \sqrt{s^2} \)
- Interquartile range: \( IQR = \text{quantile}(X,.75) - \text{quantile}(X,.25) \)
- Median absolute deviation: \( MAD = \text{median}\{|X - m|\} \), \( m = \text{median}(X) \)
Robustness

• IQR and MAD are *robust* measures of variability. Insensitive to a few outliers.

• Standard deviation is not robust. One extreme outlier can change its value drastically.

• All are *scale* parameters or statistics. When the scale of measurement is changed, they change in the same way.
Variability of Categorical Variables
Multinomial Distributions

• A categorical variable has $m$ possible values, with probabilities $p\downarrow 1, \cdots, p\downarrow m$, positive and summing to 1.

• Replicate the experiment $N$ times independently. Possibly $N=1$.

• $Y\downarrow i =$ number of occurrences of $i^{th}$ outcome.

• This is a *multinomial experiment* and the random vector $Y=(Y\downarrow 1, \cdots, Y\downarrow m)$ has a multinomial distribution.
Gini Measure of Variability

- In the multinomial distribution, each component $Y_{\downarrow i}$ has a binomial distribution with variance $Np_{\downarrow i} (1 - p_{\downarrow i})$.
- $Gini = N\sum_{1}^{m} p_{\downarrow i} (1 - p_{\downarrow i}) = N (1 - \sum_{1}^{m} p_{\downarrow i}^2)$
- Since $\sum_{1}^{m} p_{\downarrow i} = 1$, $Gini$ is maximum when each $p_{\downarrow i} = 1/m$, i.e., all category levels are equally likely, and 0 when some $p_{\downarrow i} = 1$, others = 0.
- Note: The maximum value increases with $m$.
- With data, replace $p_{\downarrow i}$ by its estimate $Y_{\downarrow i} / N$.
Entropy Measure of Variability

• $H = -N \sum 1^m p^i \log p^i$
• By continuity, define $0 \log 0 = 0$. Then $0 \leq H \leq N \log m$.
• $H = 0$ when some $p^i = 1$. $H = M \log m$ when all $p^i = 1/m$. The maximum value increases with $m$.
• With data, replace $p^i$ by its estimate $Y^i/N$. Then $H$ is related to the likelihood ratio statistic for the null hypothesis of equally likely category levels.
Correlation

To What Extent Are Variables Related?
Theoretical Covariance and Correlation

Pearson Correlation

• $X, Y$ jointly distributed random variables
• Means $\mu \downarrow x, \mu \downarrow y$, standard deviations $\sigma \downarrow x > 0$, $\sigma \downarrow y > 0$.
• $\text{cov}(X, Y) = E[(X - \mu \downarrow x)(Y - \mu \downarrow y)]$
• $\text{cor}(X, Y) = \rho \downarrow xy = \text{cov}(X, Y)/\sigma \downarrow x \sigma \downarrow y$
• $|\rho| \leq 1$, with equality iff $aX + bY = c$ for constants $a, b, c$. 
Sample Covariance and Correlation

• $s_{xy} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})(y_i - \bar{y})$
• $s_{x}^{2} = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - x)^2$
• $s_{y}^{2} = \frac{1}{n-1} \sum_{i=1}^{n} (y_i - y)^2$
• $r_{xy} = \frac{s_{xy}}{s_{x} s_{y}}$
• Random variables. $|r_{xy}| \leq 1$ with equality iff $ax_i + by_i = c$ for all $i$. 
Guess $\rho$, Guess $r$
Spearman’s Rho

• Given data $(x\downarrow 1, y\downarrow 1), \cdots (x\downarrow n, y\downarrow n)$, rank the $x$'s and also rank the $y$'s. Let $u\downarrow i = \text{rank}(x\downarrow i)$ and $v\downarrow i = \text{rank}(y\downarrow i)$.

• Then calculate the Pearson correlation of the pairs $(u\downarrow 1, v\downarrow 1), \cdots , (u\downarrow n, v\downarrow n)$.

• This is Spearman’s rho $\rho \downarrow s$.

• If $X$ and $Y$ are independent, the distribution of $\rho \downarrow s$ does not depend on their distributions.

• Provides a nonparametric or distribution-free test of no association between $X$ and $Y$. 
Kendall’s Tau

• Count the number \( c \) of pairs of indices \((i, j)\) with \( i < j \) and \((x_{\downarrow i} - x_{\downarrow j})(y_{\downarrow i} - y_{\downarrow j}) > 0\). These are concordant pairs.

• The number \( d \) of discordant pairs is \( p - c \), where \( p = \frac{1}{2} n(n - 1) \).

• \( \tau = c - d/p \)

• \( \tau \) is distribution free if \( X \) and \( Y \) are independent.
Comparison

\[ r = 0.86, \quad r = 0.65, \quad r = 0.50 \]
Variance-Covariance Matrices

Random Vectors
Variance-Covariance Matrix

- \( X \downarrow 1, X \downarrow 2, \ldots, X \downarrow m \) jointly distributed numeric variables.
- \( X = (X \downarrow 1, X \downarrow 2, \ldots, X \downarrow m)^{\top} \in \mathbb{R}^{m \times 1} \) is a random vector.
- \( V = V(X) = (v \downarrow ij) \in \mathbb{R}^{m \times m} \), where \( v \downarrow ij = \text{cov}(X \downarrow i, X \downarrow j) = \rho \downarrow ij \sigma_i \sigma_j \), \( \rho \downarrow ij = \text{cor}(X \downarrow i, X \downarrow j) \).
- Positive definite, symmetric matrix with positive eigenvalues, orthogonal eigenvectors.
- Given \( n \) sample observations of \( X \), the sample variance-covariance matrix \( V \) has sample correlations and standard deviations.
Principal Components

• $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m > 0$ the ordered eigenvalues of $V$.
• $u_1, u_2, \ldots, u_m$ corresponding orthogonal unit eigenvectors.
• $u_1 \cdot X, u_2 \cdot X, \ldots, u_m \cdot X$ are uncorrelated. Called the principal components of the random vector $X$.
• $\lambda_1 = \text{var}(u_1 \cdot X), \lambda_2 = \text{var}(u_2 \cdot X), \text{etc.}$
Importance of Principal Components

- $\sum_{i=1}^{m} \lambda_i = \sum_{i=1}^{m} \text{var}(X_i)$
- If the first few largest $\lambda_i$ strongly dominate, most of the variation of the random vector $X$ is captured by the first few principal components.
- Useful as a dimensionality reduction tool.
Wing shape is quantified by noting the location of landmarks defined by the intersection of veins with each other or the wing margin.
Total of 15 landmarks used.
Some of the Variables
Variances of Principal Components

[1] 0.0922 0.0405 0.0116 0.0054 0.0004 0.0002 0.0001 0.0001 0.0000 0.0000
[11] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000
[21] 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000 0.0000

Total variation is 0.15. Top three carry most of it.
Principal Components by Species
Classification Trees

• Splitting of nodes always decreases Gini or entropy. So splits always increase “purity” of terminal nodes.

• Split nodes on single variables, nodes and variables chosen to maximize the decrease in total Gini or entropy.

• Stop when the decrease falls below a threshold or when nodes get too small.
Example