What is correlation (what does it mean for data to be correlated)?

- A measure of the strength and direction of a linear relationship between two random variables
- In symbols:
  \[ \rho_{XY} = \frac{E[(X - \mu_X)(Y - \mu_Y)]}{\sqrt{\text{Var}(X) \text{Var}(Y)}} = \frac{\text{Cov}(X, Y)}{\sigma_X \sigma_Y} \]

  2 “Standardized covariance”
  8 “Function of the angle between the two variable vectors”
  10 “Estimation from the balloon rule”

Example: Measuring blood pressure

Where have you seen correlation before?

- Can be used as a bivariate descriptive statistic (e.g., relationship between Y and X)
- Generally not of scientific interest
- Will tend to increase (magnitude) as:
  - the magnitude of the slope of the line increases
  - the variance of Y|X decreases
  - the variance of X increases
  increasing the sample size need not “improve” correlation
Correlation for this course will focus on that between more than one measurement of an independent variable (e.g., $X_i^1$ and $X_i^2$)

- Still generally not of scientific interest: "nuisance parameter"

How does correlation happen?

- Cluster or multi-level sampling
  - sampling within institutions (e.g., schools, hospitals)
  - sampling within families
  - complex survey samples

- Multivariate outcomes
  - cholesterol and blood pressure considered simultaneously as an outcome

- Repeated measurements on the same subject
  - necessarily over time
  - varying experimental conditions
  - prolonged exposure

Motivation for longitudinal studies

- Convenience of existing study population
- Efficiency of using subjects as own comparison
- Scientific question of interest is inherently longitudinal in nature
  - effects over time
  - within subject

Convenience

- Taking multiple measurements on each individual is easier than recruiting additional subjects
- E.g., Tumor reoccurrences
  - natural disease course is to have multiple reoccurrences
  - interest is in reducing reoccurrences (multiple ways to interpret "reduced reoccurrences")
Efficiency
▶ Efficiency of using subjects as their own comparison (e.g., own “control”)
  ▶ the primary comparison of interest within subjects may have less variability
    ▶ less variable statistic ⇒ detect smaller effects/more power/smaller sample size required
    ▶ E.g., adjusting for baseline measurements
    ▶ E.g., cross-over study of new treatment

Example: Independent Samples
▶ Large between-subject variability will inhibit our ability to detect difference between groups
  ▶ standard errors are generally proportional to within group standard deviation divided by $\sqrt{n}$

Example: Cross-over
▶ (high) Correlation between measurements taken on the same individual can increase the precision

Longitudinal Scientific Questions of Interest
▶ Scientific questions about effects that occur over time
  ▶ studies designed to examine population time trends in response
    ▶ E.g., rate (slope) of progression of retinopathy in a population of diabetic patients over time
    ▶ E.g., time to relapse in cancer study
Example Time trends in Measurements
▶ Mean differences in mean hemoglobin A1C among diabetic patients
▶ may have trends in both means and variability

Longitudinal Scientific Questions of Interest
▶ Trends for the average measurements in the population may not represent trends of specific individuals
▶ response over time may be restricted to subgroups of subjects
▶ response over time may be transient

Longitudinal Scientific Questions of Interest
▶ Scientific questions may instead be regarding effects that occur within subjects
▶ distribution of rates (slopes) of progression of retinopathy in population over time
▶ effect of varying risk factors within individuals

May have different trajectory over time in subgroup
May all have an effect, but at different times (and magnitudes)

Measures of outcome
- Scientific hypotheses are addressed statistically by comparing summary measures for a distribution
  - mean response in population
  - median response in population
  - percentage of individuals exceeding some threshold
  - etc.

Choice of summary measure
- The choice of summary measure should be decided by (in order of importance):
  - greatest clinical (scientific) relevance
  - something the treatment is expected to effect
  - measured reliably
  - statistical precision
  - E.g.,
    - scientific relevance may be associated with achieving some threshold
    - mean can be sensitive to outliers
    - relative precision of mean versus median depends on presence of outliers

Choice of summary measure
- For longitudinal studies, individuals may have multiple measurements taken over time
  - How one may define an individual response of interest can vary
    - response at fixed time
    - response at multiple fixed times
    - average response over time (area under the curve, “time-weighted average”)
    - rate of change in response
    - time to reaching threshold of interest
    - number of times reaching threshold of interest
Choice of summary measure
▶ For longitudinal studies, choice of response should reflect scientific relevance, plausibility of effect, reliably measured, precision
▶ final measurement may be more important than any effects observed earlier
▶ summarizing response at multiple time points reflects population rather than individuals
▶ weight average over time sensitive to potential transient effects
▶ differences in time to event may be meaningless, fact that the event occurred is key

Why is it of note? Why can’t we ignore?
▶ Consider the impact on:
  ▶ means
  ▶ rates (prevalence, incidence)
  ▶ regression coefficients
  ▶ variance estimates (precision)

Means
▶ No impact in the balanced case:
  ▶ Let $Y_{ij}$ be the outcome of interest on the $j^{th}$ unit from the $i^{th}$ cluster where $I$ is the total number of clusters. Suppose all clusters are the same size ($n_j = n$), that all observations within a cluster have the same mean $\mu_i$, and $\text{Var}(Y_{ij}) = \sigma^2$ for all $i$, $j$. Let $\rho$ denote the common correlation between observations within a cluster.
  ▶ The sample mean for the $i^{th}$ cluster is $\bar{Y}_i = \frac{1}{n} \sum_{j=1}^{n} Y_{ij}$
  ▶ Then the expected value of the sample mean is $E(\bar{Y}_i) = E(\sum_{j=1}^{n} Y_{ij}) = \frac{1}{n} \sum_{j=1}^{n} E(Y_{ij}) = \mu_i$
  (although when unbalanced, point estimates may be off for population parameters - more (less) weight can be placed on some subjects compared to others)
  ▶ What about rates and regression coefficients?

Rates and regression coefficients
▶ Rates or regression coefficients are weighted means - same as previous.
▶ What about estimates of precision?
  ▶ YES: consider $\text{Var}(\bar{Y}_i)$
Properties of Variance

- Variance of sums (differences) of random variables \( X \) and \( Y \)
  \[
  \text{Var}(X + Y) = \text{Var}(X) + \text{Var}(Y) + 2\text{Cov}(X, Y)
  \]
  \[
  \text{Var}(X - Y) = \text{Var}(X) + \text{Var}(Y) - 2\text{Cov}(X, Y)
  \]

As such, we can calculate the variability of sample means computed across replicate experiments

- Suppose there are \( n \) independent measurements sampled from a population wherein
  \( \mu \) the population average is \( \mu \)
  \( \sigma^2 \) the variance of measurements in the population is \( \sigma^2 \)
  Then the average sample mean will be \( \mu \) and the variance of the sample mean will be \( \sigma^2/n \)
  \[
  \text{Var}(\bar{X}) = \text{Var} \left( \frac{1}{n} \sum_{i=1}^{n} X_i \right) = \frac{1}{n^2} \sum_{i=1}^{n} \text{Var}(X_i) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}
  \]

- Suppose instead there are \( n \) correlated measurements sampled from a population wherein
  \( \mu \) and \( \sigma^2 \) (effectively just as before) with \( i \) and \( j \) from 1 to \( m \) individuals each with \( k \) measurements respectively
  \( \rho \) measurements on the same individual are correlated
  Sample mean
  \[
  \bar{X} = \frac{1}{mk} \sum_{i=1}^{m} \sum_{j=1}^{k} X_{ij}
  \]
  \[
  \text{Var}(\bar{X}) = \frac{\sigma^2}{mk} (1 + \rho(k - 1))
  \]

- Averages of correlated observations are less precise
- Differences between correlated observations are more precise

Implications of impact on precision

- Inference can be (seriously) incorrect
- Ultimately interested in four numbers
  - ‘point’ estimate of effect
  - lower and upper bound of an ‘interval’ estimate
  - quantification of strength of evidence based on data available for (against) particular ‘hypotheses’

Note: This is regardless of frequentist or bayesian mindsets.
- Poorly estimated precision affects three of the four numbers - generally:
  - confidence intervals can be too narrow or too wide depending on the comparison performed, hence will not be an accurate representation of our true level of confidence
  - p-values too small or too big, i.e., wrong
- Also have inefficient estimates.
How to “handle” the correlation
- Reduce measurements on a given cluster to a single measurement and proceed
  - can be very difficult if any missing data (generally always occurs to some degree)
  - cannot evaluate interaction
  - may not be most efficient statistically
- Estimate correlation within clusters and adjust standard errors for population based models
  - GEE, marginal models
  - “Robust” sandwich estimation for variance
- Adjust estimates of means for taking correlated nature into account via ‘random effects’
  - “Mixed effects models”

Examples
- Beta-carotene study
  - clinical trial of beta-carotene supplements on plasma levels of beta-carotene and vitamin E
  - subjects randomized to five different dose groups
  - measurements at baseline, 3mos, 9mos, and 3mos post terminating treatment
  - What differences are there in changes of plasma level beta-carotene over time between dose groups?

- Somatosensory evoked potential (SEP)
  - measurements on a sample of healthy adults
  - nerve conduction times are measured
  - four separate measurements for each leg

- Six Cities Study
  - prospective observational study of the effects of air pollution in adults and children
  - six U.S. cities, years 1, 4, and 7
  - outcome = forced expiratory volume in 1 second (FEV-1)
  - predictors = age, height, weight, sex, pack-years, cigarettes/week, fume exposure, dust exposure, bronchitis, emphysema, pneumonia
  - not all people have observations at all years
Women’s Fungal Study (CPCA 010)
- prospective randomized clinical trial on primary and secondary prophylaxis on mucosal candidiasis
- HIV+ women with a CD4+ cell count ever ≤ 300 cells/mm³
- 14 clinical units across the U.S.
- 29 median months of follow-up
- treatment = fluconazole vs. placebo
- outcome = positive culture for vaginal candidiasis
- predictors = age, race, IV drug use, anti-retroviral drug use, history of candidiasis, progression of disease, CD4+ cell count
Meta-analysis of pre-eclampsia trials
- nine randomized trials comparing a diuretic to a control treatment for pre-eclampsia
- outcome = pre-eclampsia (yes/no)
- predictors = treatment trial
Correlated outcomes can occur when:

- multiple observations are taken from a single individual across time (cluster = ?)
- multiple observations are taken from a single individual at one time (cluster = ?)
- individuals are sampled in a geographic framework (cluster = ?)
- individuals are sampled in a clustered framework

---

**Basic Notation**

\[
\begin{align*}
  y_{ij} &= \text{response value at the } j^{th} \text{ time point for the } i^{th} \text{ individual} \\
  y_i &= \text{response value for the } j^{th} \text{ individual in the } i^{th} \text{ cluster} \\
  x_{ij1} &= \text{first covariate's value} \\
  x_{ij2} &= \text{second covariate's value} \\
  \vdots \\
  x_{ijp} &= \text{p}^{th} \text{ covariate's value}
\end{align*}
\]

---

**Longitudinal or Time Series?**

<table>
<thead>
<tr>
<th></th>
<th>Individuals (things measured)</th>
<th>Time Points (times measured)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Longitudinal</td>
<td>many</td>
<td>few</td>
</tr>
<tr>
<td>Time series</td>
<td>few</td>
<td>many</td>
</tr>
</tbody>
</table>

- boundaries between the two types are not clear
- models to analyze them can be very different
- focus in longitudinal is often on comparing groups or assessing covariate effects
- focus in time series is often on prediction (forecasting) or signal extraction

---

**Models will generally look like:**

\[
\begin{align*}
  y_{ij} &= \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp} + \epsilon_{ij} \\
  E[y_{ij}] &= \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp}
\end{align*}
\]

which is the familiar form for multiple linear regression. Consider a similar form:

\[
g(E[y_{ij}]) = \beta_0 + \beta_1 x_{ij1} + \cdots + \beta_p x_{ijp}
\]

where \( g(\cdot) \) is referred to as a “link function”. This is the form of a **generalized linear model** (GLM). If \( g(\cdot) \) is the identity function, we are back to the equation at the top - a special case of GLM. Other choices of \( g(\cdot) \) are commonly used: logit and log for situations such as, e.g., binary outcomes, or if we wanted to do Poisson regression.
Our data now have two levels: 'between' and 'within'.

<table>
<thead>
<tr>
<th></th>
<th>Longitudinal</th>
<th>Clustered</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>Individual</td>
<td>Cluster</td>
</tr>
<tr>
<td>2</td>
<td>Observation within</td>
<td>Observation within</td>
</tr>
<tr>
<td></td>
<td>individual</td>
<td>cluster</td>
</tr>
</tbody>
</table>

We need to understand any patterns in the data at both levels.