# Statistical Distribution Assumptions of General Linear Models

Applied Multilevel Models for Cross-Sectional Data
Lecture 3

ICPSR Summer Workshop
University of Colorado Boulder

# Covered this Section

- The building blocks of multilevel models -- basics of mathematical statistics:
  - Random variables: definitions and types
  - Univariate distributions
    - General terminology
    - Univariate normal (aka, Gaussian)
    - Other popular (continuous) univariate distributions
  - Types of distributions: marginal, conditional, and joint
  - Expected values: means, variances, and the algebra of expectations
  - Linear combinations of random variables

- The finished product: How the GLM fits within statistics
  - The GLM with the normal distribution
  - The statistical assumptions of the GLM

# RANDOM VARIABLES AND STATISTICAL DISTRIBUTIONS

# Random Variables

**Random:** situations in which the certainty of the outcome is unknown and is at least in part due to chance

**+**

**Variable:** a value that may change given the scope of a given problem or set of operations

**=**

**Random Variable:** a variable whose outcome depends on chance

(possible values might represent the possible outcomes of a yet-to-be-performed experiment)

Today we will denote a random variable with a lower-cased: $x$

# Types of Random Variables

- Random variables have different types:

1. <u>Continuous</u>
   - Examples of continuous random variables:
     - $x$ represents the height of a person, drawn at random
     - $Y_p$ (the outcome/DV in a GLM)

2. <u>Discrete (also called categorical, generally)</u>
   - Example of discrete:
     - $x$ represents the gender of a person, drawn at random

3. <u>Mixture of Continuous and Discrete:</u>
   - Example of mixture:
     - $x$ represents $\begin{cases} \text{response time (if between 0 and 45 seconds)} \\ 0 \end{cases}$

# Key Features of Random Variables

- Random variables each are described by a **probability density/mass function (PDF)** $f(x)$ that indicates relative frequency of occurrence
  - A PDF is a mathematical function that gives a rough picture of the distribution from which a random variable is drawn

- The type of random variable dictates the name and nature of these functions:
  - Continuous random variables:
    - $f(x)$ is called a probability density function
    - Area under curve must equal 1 (found by calculus – integration)
    - Height of curve (the function value $f(x)$):
      - Can be any positive number
      - Reflects relative likelihood of an observation occurring

# Uses of Distributions in Data Analysis

- Statistical models make distributional assumptions on various parameters and/or parts of data

- These assumptions govern:
  - How models are estimated
  - How inferences are made
  - How missing data may be imputed

- If data do not follow an assumed distribution, inferences may be inaccurate
  - Sometimes a very big problem, other times not so much

# CONTINUOUS UNIVARIATE DISTRIBUTIONS

# Continuous Univariate Distributions

- To demonstrate how continuous distributions work and look, we will discuss two:
  - Uniform distribution
  - Normal distribution
- Each are described a set of **parameters**, which we will later see are what give us our inferences when we analyze data
- What we then do is put constraints on those parameters based on hypothesized effects in data

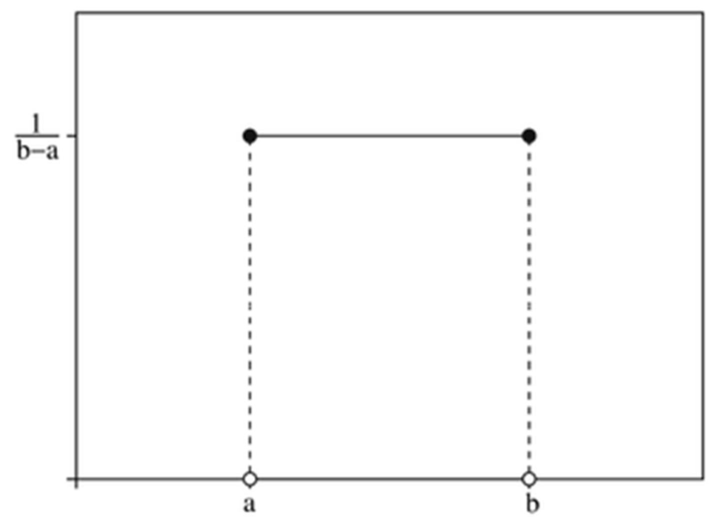
# Uniform Distribution

- The uniform distribution is shown to help set up how continuous distributions work

- For a continuous random variable $x$ that ranges from $(a, b)$, the uniform probability density function is:

$$f(x) = \frac{1}{b - a}$$

- The uniform distribution has two parameters:
  - $a$ – the lower limit
  - $b$ – the upper limit




- $x \sim U(a, b)$

# More on the Uniform Distribution

- To demonstrate how PDFs work, we will try a few values:

| $\boldsymbol{x}$ | $\boldsymbol{a}$ | $\boldsymbol{b}$ | $\boldsymbol{f(x)}$ |
|---|---|---|---|
| .5 | 0 | 1 | $\frac{1}{1-0}=1$ |
| .75 | 0 | 1 | $\frac{1}{1-0}=1$ |
| 15 | 0 | 20 | $\frac{1}{20-0}=.05$ |
| 15 | 10 | 20 | $\frac{1}{20-10}=.1$ |

- The uniform PDF has the feature that all values of $x$ are **equally likely** across the sample space of the distribution
  - Therefore, you do not see $x$ in the PDF $f(x)$

- The mean of the uniform distribution is $\frac{1}{2}(a+b)$
- The variance of the uniform distribution is $\frac{1}{12}(b-a)^2$

# Univariate Normal Distribution

- For a continuous random variable $x$ (ranging from $-\infty$ to $\infty$) the univariate normal distribution function is:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma_x^2}} \exp\left(-\frac{(x-\mu_x)^2}{2\sigma_x^2}\right)$$

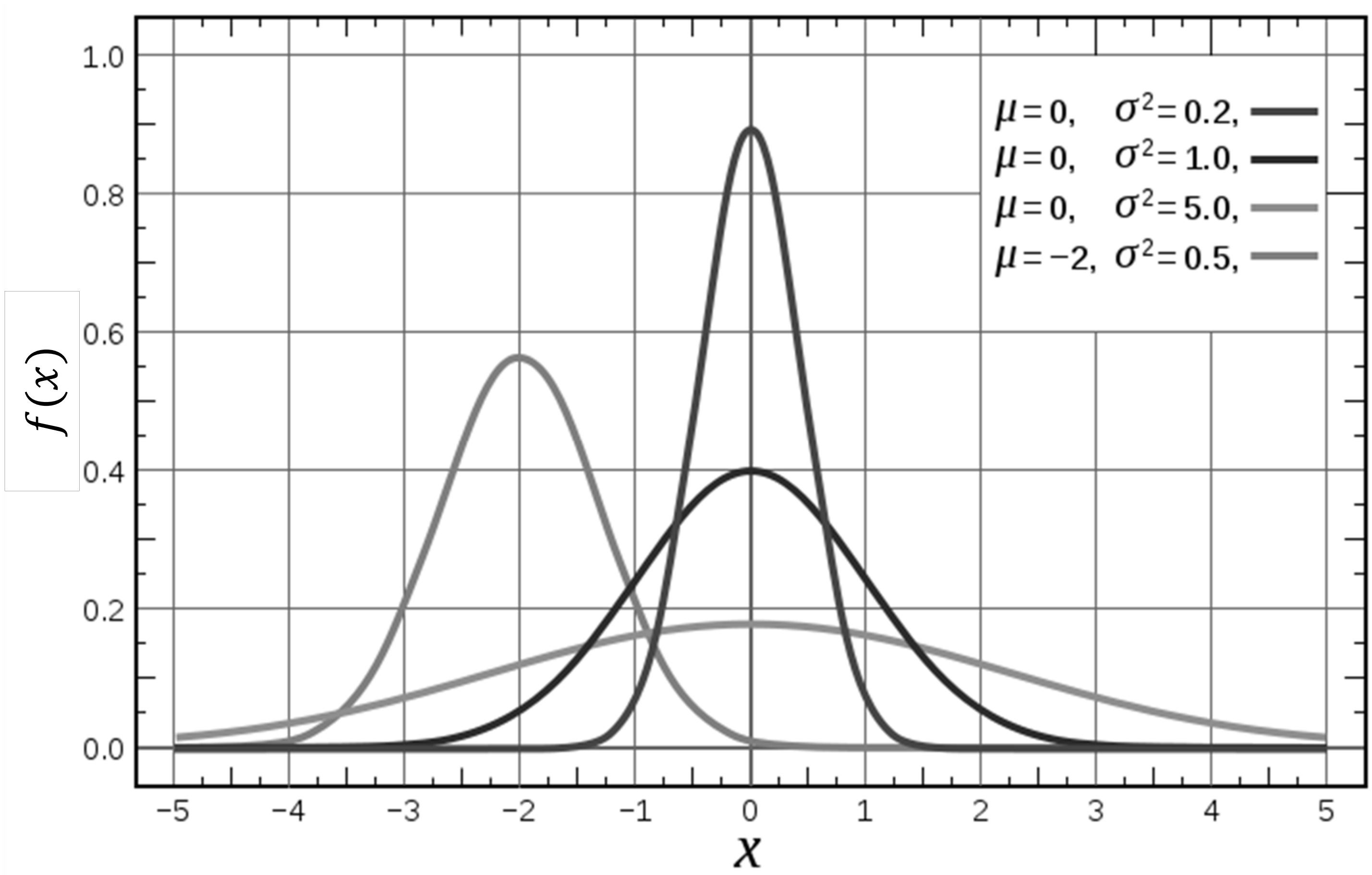
- The shape of the distribution is governed by two parameters:
  - The mean $\mu_x$
  - The variance $\sigma_x^2$
  - These parameters are called **sufficient statistics** (they contain all the information about the distribution)

- The skewness (lean) and kurtosis (peakedness) are fixed

- Standard notation for normal distributions is $x \sim N(\mu_x, \sigma_x^2)$
  - Read as: "$x$ follows a normal distribution with a mean $\mu_x$ and a variance $\sigma_x^2$"

- Linear combinations of random variables following normal distributions result in a random variable that is normally distributed
  - You'll see this later with respect to the GLM…

# Univariate Normal Distribution




$f(x)$ gives the height of the curve (relative frequency) for any value of $x$, $\mu_x$, and $\sigma_x^2$

# More of the Univariate Normal Distribution

- To demonstrate how the normal distribution works, we will try a few values:

| $\boldsymbol{x}$ | $\boldsymbol{\mu_x}$ | $\boldsymbol{\sigma_x^2}$ | $\boldsymbol{f(x)}$ |
|---|---|---|---|
| .5 | 0 | 1 | 0.352 |
| .75 | 0 | 1 | 0.301 |
| .5 | 0 | 5 | 0.079 |
| .75 | -2 | 1 | 0.009 |
| -2 | -2 | 1 | 0.399 |

- The values from $f(x)$ were obtained by using Excel
  - The "=normdist()" function
  - Most statistics packages have a normal distribution function

- The mean of the normal distribution is $\mu_x$
- The variance of the normal distribution is $\sigma_x^2$

# MARGINAL, JOINT,
# AND CONDITIONAL DISTRIBUTIONS

# Moving from One to Multiple Random Variables

- When more than one random variable is present, there are several different types of statistical distributions:

- We will first consider two discrete random variables:
  - $x$ is the outcome of the flip of a penny $(H_p, T_p)$
    - $f(x = H_p) = .5$ ; $f(x = T_p) = .5$
  - $z$ is the outcome of the flip of a dime $(H_d, T_d)$
    - $f(z = H_d) = .5$ ; $f(z = T_d) = .5$

- We will consider the following distributions:
  - Marginal distribution
    - The distribution of one variable only (either $f(x)$ ***<u>or</u>*** $f(z)$)
  - Joint distribution
    - $f(x, z)$: the distribution of both variables (both $x$ ***<u>and</u>*** $z$)
  - Conditional distribution
    - The distribution of one variable, conditional on values of the other:
      - $f(x|z)$: the distribution of $x$ given $z$
      - $f(z|x)$: the distribution of $z$ given $x$

# Marginal Distributions

- Marginal distributions are what we have worked with exclusively up to this point: they represent the distribution of one variable by itself
  - Continuous univariate distributions:
    - Uniform
    - Normal
  - Categorical distributions in our example:
    - The flip of a penny $f(x)$
    - The flip of a dime $f(z)$

# Joint Distributions

- Joint distributions describe the distribution of more than one variable simultaneously
  - Representations of multiple variables collected
  - Multilevel data are multivariate – the level-1 observations are like multiple observations from a level-2 sampling unit

- Commonly, the joint distribution function is denoted with all random variables separated by commas
  - In our example, $f(x, z)$ is the joint distribution of the outcome of flipping both a penny and a dime
    - As both are discrete, the joint distribution has four possible values: $f(x = H_p, z = H_d), f(x = H_p, z = T_d), f(x = T_p, z = H_d), f(x = T_p, z = T_d)$

- Joint distributions are **multivariate distributions**
  - We will cover the multivariate normal distribution as it is the engine that drives most multilevel analyses
  - For our purposes, we will use joint distributions to introduce joint likelihoods – used in maximum likelihood estimation

# Joint Distributions of Independent Random Variables

- Random variables are said to be independent if the occurrence of one event makes it neither more nor less probable of another event
  - For joint distributions, this means: $f(x,z) = f(x)f(z)$

- In our example, flipping a penny and flipping a dime are independent – so we can complete the following table of their joint distribution:

Joint
(Penny, Dime)

Dime

| | $\boldsymbol{z = H_d}$ | $\boldsymbol{z = T_d}$ | |
|---|---|---|---|
| $x = H_p$ | $f(x = H_p, z = H_d)$ | $f(x = H_p, z = T_d)$ | $f(x = H_p)$ |
| $x = T_p$ | $f(x = T_p, z = H_d)$ | $f(x = T_p, z = T_d)$ | $f(x = T_d)$ |
| | $f(z = H_d)$ | $f(z = T_d)$ | |

Penny

Marginal
(Penny)

Marginal
(Dime)

# Joint Distributions of Independent Random Variables

- Because the coin flips are independent, this becomes:

Joint (Penny, Dime)

Dime

Penny

| | $z = H_d$ | $z = T_d$ | |
|---|---|---|---|
| $x = H_p$ | $f(x = H_p)f(z = H_d)$ | $f(x = H_p)f(z = T_d)$ | $f(x = H_p)$ |
| $x = T_p$ | $f(x = T_p)f(z = H_d)$ | $f(x = T_p)f(z = T_d)$ | $f(x = T_d)$ |
| | $f(z = H_d)$ | $f(z = T_d)$ | |

Marginal (Penny)

Marginal (Dime)

- Then, with numbers:

Joint (Penny, Dime)

Dime

Penny

| | $z = H_d$ | $z = T_d$ | |
|---|---|---|---|
| $x = H_p$ | .25 | .25 | .5 |
| $x = T_p$ | .25 | .25 | .5 |
| | .5 | .5 | |

Marginal (Penny)

Marginal (Dime)

# Marginalizing Across a Joint Distribution

- If you had a joint distribution, $f(x,z)$, but wanted the marginal distribution of either variable ($f(x)$ or $f(z)$) you would have to **marginalize** across one dimension of the joint distribution

- For categorical random variables, **marginalize = sum across**

$$f(x) = \sum_{z} f(x,z)$$

  - For example $f(x = H_p) = f(x = H_p, z = H_p) + f(x = H_p, z = T_p) = .5$

- For continuous random variables, **marginalize = integrate across**
  - No integration needed from you – just a conceptual understanding
  - Here, the integral = an eraser!

$$f(x) = \int_{z} f(x,z)dz$$

# Conditional Distributions

- For two random variables $x$ and $z$, a conditional distribution is written as: $f(z|x)$
  - The distribution of $z$ given $x$

- The conditional distribution is also equal to the joint distribution divided by the marginal distribution of the conditioning random variable

$$f(z|x) = \frac{f(z,x)}{f(x)}$$

- Conditional distributions are found everywhere in statistics
  - As we will see, the general linear model uses the conditional distribution of the dependent variable (where the independent variables are the conditioning variables)

# Conditional Distributions

- For discrete random variables, the conditional distribution is easy to show:

Dime

Joint (Penny, Dime)

Penny

| | $\boldsymbol{z = H_d}$ | $\boldsymbol{z = T_d}$ | |
|---|---|---|---|
| $x = H_p$ | .25 | .25 | .5 |
| $x = T_p$ | .25 | .25 | .5 |
| | .5 | .5 | |

Marginal (Penny)

Marginal (Dime)

**Conditional:** $\boldsymbol{f(z|x = H_p)}$:

$$f(z = H_d|x = H_p) = \frac{f(z = H_d, x = H_p)}{f(x = H_p)} = \frac{.25}{.5} = .5$$

$$f(z = T_d|x = H_p) = \frac{f(z = T_d, x = H_p)}{f(x = H_p)} = \frac{.25}{.5} = .5$$

We will show a continuous conditional distribution with the GLM in a few slides

# EXPECTED VALUES AND THE ALGEBRA OF EXPECTATIONS

# Expected Values

- Expected values are statistics taken the sample space of a random variable: they are essentially weighted averages
- The weights used in computing this average correspond to the probabilities (for a discrete random variable) or to the densities (for a continuous random variable)
- Notation: the expected value is represented by: $E(x)$
  - *The actual statistic that is being weighted by the PDF is put into the parentheses where $x$ is now*
- Expected values allow us to understand what a statistical model implies about data:
  - How a GLM specifies the (conditional) mean and variance of a DV

# Expected Value Calculation

- For discrete random variables, the expected value is found by:

$$E(x) = \sum_{x} xP(X = x)$$

- For example, the expected value of a roll of a die is:

$$E(x) = (1)\frac{1}{6} + (2)\frac{1}{6} + (3)\frac{1}{6} + (4)\frac{1}{6} + (5)\frac{1}{6} + (6)\frac{1}{6} = 3.5$$

- For continuous random variables, the expected value is found by:

$$E(x) = \int_{x} xf(x)dx$$

- We won't be calculating theoretical expected values with calculus…we use them only to see how models imply things about our data

# Variance and Covariance…As Expected Values

- A distribution's theoretical variance can also be written as an expected value:

$$V(x) = E\left(x - E(x)\right)^2 = E(x - \mu_x)^2$$

  - This formula will help us understand predictions made GLMs and how that corresponds to statistical parameters we interpret

- For a roll of a die, the theoretical variance is:

$$V(x) = E(x - 3.5)^2 = \frac{1}{6}(1 - 3.5)^2 + \frac{1}{6}(2 - 3.5)^2 + \frac{1}{6}(3 - 3.5)^2 + \frac{1}{6}(4 - 3.5)^2 + \frac{1}{6}(5 - 3.5)^2 + \frac{1}{6}(6 - 3.5)^2 = 2.92$$

  - Likewise, the SD is then $\sqrt{2.92} = 1.71$

- Likewise, for a pair of random variables $x$ and $z$, the covariance can be found from their joint distributions:

$$Cov(x, z) = E(xz) - E(x)E(z) = E(xz) - \mu_x\mu_z$$

# THE GENERAL LINEAR MODEL WITH WHAT WE HAVE LEARNED TODAY

# The General Linear Model, Revisited

- The general linear model for predicting Y from X and Z:

$$Y_p = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p$$

In terms of random variables, under the GLM:

- $e_p$ is considered random: $e_p \sim N(0, \sigma_e^2)$

- $Y_p$ is dependent on the linear combination of $X_p, Z_p$, and $e_p$

- The GLM provides a model for the **conditional distribution** of the dependent variable, where the conditioning variables are the independent variables: $f(Y_p | X_p, Z_p)$
  - There are no assumptions made about $X_p$ and $Z_p$ - they are constants
  - The regression slopes $\beta_0, \beta_1, \beta_2, \beta_3$ are constants that are said to be fixed at their values (hence, called fixed effects)

# Combining the GLM with Expectations

- Using the algebra of expectations predicting Y from X and Z:

**The expected value (mean) of $\boldsymbol{f\left(Y_p \middle| X_p, Z_p\right)}$:**

$$\hat{Y}_p = E\left(Y_p\right) = E\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p\right)$$

Constants

Random Variable with $E\left(e_p\right) = 0$

$$= \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + E\left(e_p\right) = \beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p$$

**The variance of $\boldsymbol{f\left(Y_p \middle| X_p, Z_p\right)}$:**

$$V\left(Y_p\right) = V\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p + e_p\right) = V\left(e_p\right) = \sigma_e^2$$

# Distribution of $f\left(Y_p \middle| X_p, Z_p\right)$

- We just found the mean (expected value) and variance implied by the GLM for the conditional distribution of $Y_p$ given $X_p$ and $Z_p$
- The next question: what is the distribution of $f\left(Y_p \middle| X_p, Z_p\right)$?
- Linear combinations of random variables that are normally distributed result in variables that are normally distributed
- Because $e_p \sim N(0, \sigma_e^2)$ is the only random term in the GLM, the resulting conditional distribution of $Y_p$ is normally distributed:

$$Y_p | X_p, Z_p \sim N\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p, \sigma_e^2\right)$$

Model for the means: from fixed effects; literally gives mean of $f\left(Y_p \middle| X_p, Z_p\right)$

Model for the variances: from random effects; gives variance of $f\left(Y_p \middle| X_p, Z_p\right)$

# Examining What This Means in the Context of Data

- If you recall from our first lecture, the final model we decided to interpret:

- Model 5

$$W_p = \beta_0 + \beta_1\left(H_p - \overline{H}\right) + \beta_2 F_p + \beta_3\left(H_p - \overline{H}\right)F_p + e_p$$

where $e_p \sim N(0, \sigma_e^2)$

- From SAS:

| Parameter | Estimate | Standard Error | t Value | Pr > \|t\| |
|---|---|---|---|---|
| Intercept | 222.1841719 | 0.83809108 | 265.11 | <.0001 |
| heightMC | 3.1897275 | 0.11135027 | 28.65 | <.0001 |
| female | -82.2719216 | 1.21109969 | -67.93 | <.0001 |
| heightMC*female | -1.0938553 | 0.16777741 | -6.52 | <.0001 |

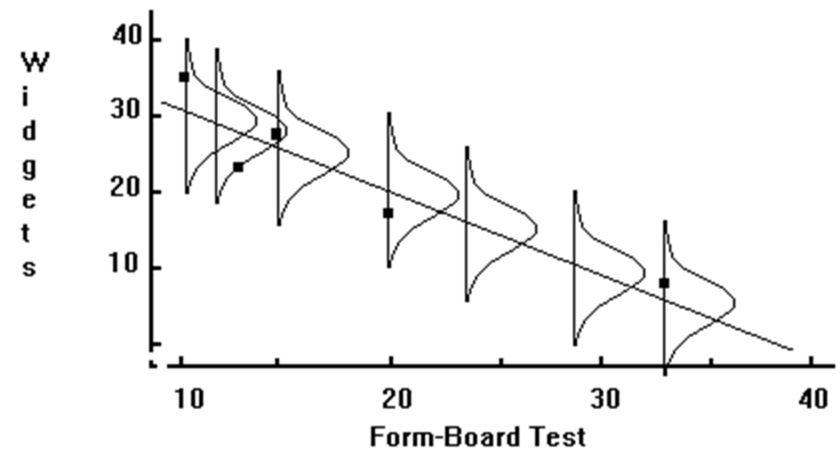
# Picturing the GLM with Distributions

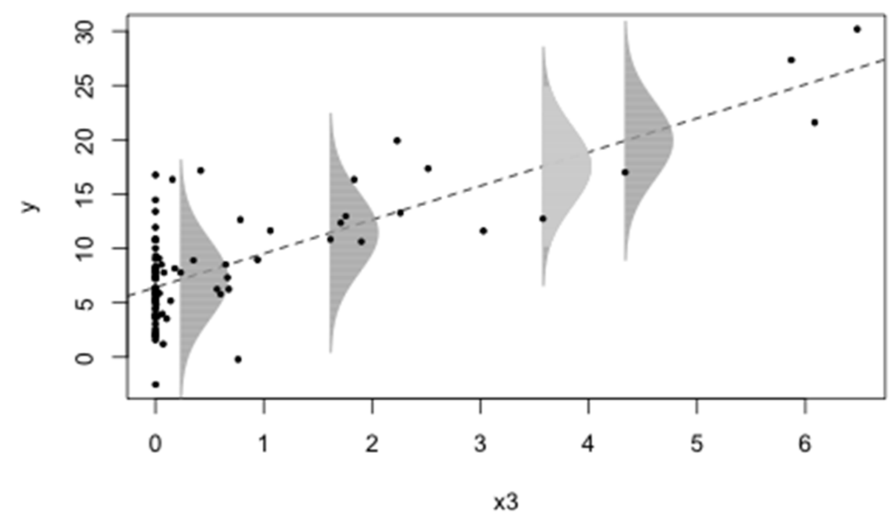
The distributional assumptions of the GLM are the reason why we do not need to worry if our dependent variable is normally distributed

Our dependent variable should be **conditionally** normal

We can check this assumption by checking our assumption about the residuals,

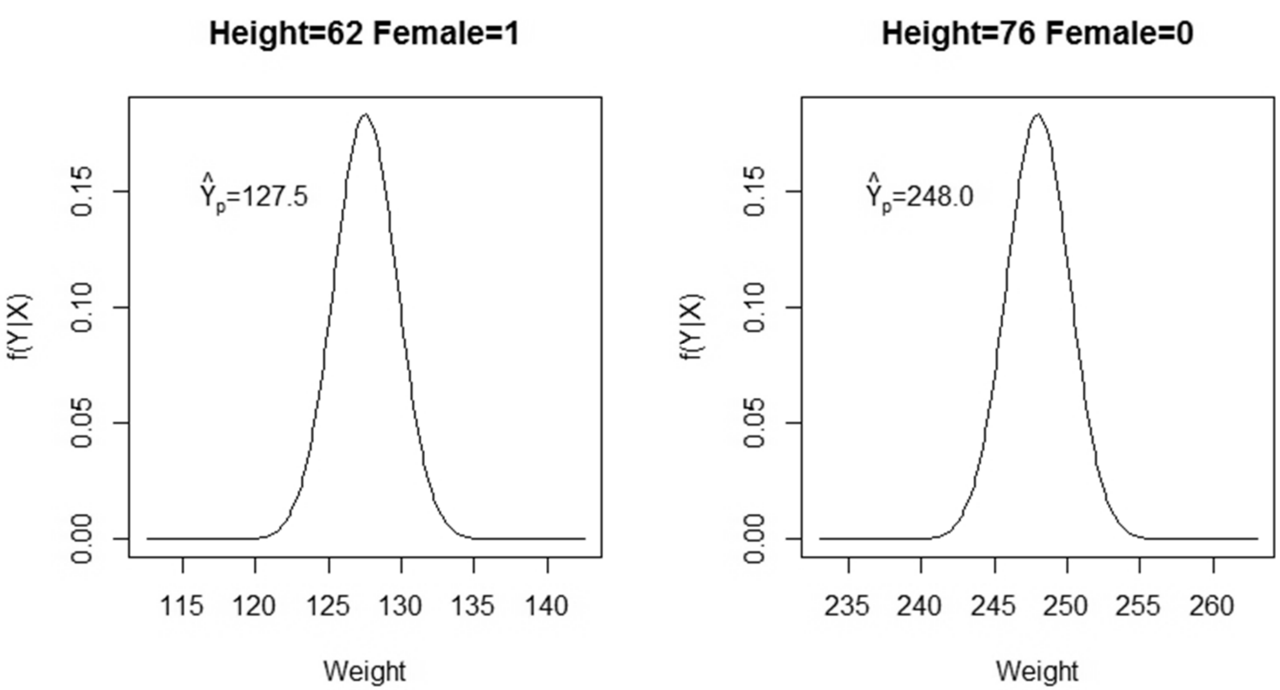
$$e_p \sim N(0, \sigma_e^2)$$



# More Pictures of the GLM

- Treating our estimated values of the slopes $(\beta_0, \beta_1, \beta_2, \beta_3)$ and the residual variance $(\sigma_e^2)$ as the true values* we can now see what the theoretical* distribution of $f\left(Weight_p \middle| Height_p, Female_p\right)$ looks like for a given set of predictors




*Note: these distributions change when sample estimates are used (think standard error of the prediction)

# Behind the Pictures…

- To emphasize the point that PDFs provide the height of the line, here is the normal PDF (with numbers) that produced those plots:

$$f\left(W_p \middle| H_p, F_p\right) = \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(W_p - \widehat{W}_p\right)^2}{2\sigma_e^2}\right)$$

**Model for the Means**

$$= \frac{1}{\sqrt{2\pi\sigma_e^2}} \exp\left(-\frac{\left(W_p - \left(\beta_0 + \beta_1\left(H_p - \bar{H}\right) + \beta_2 F_p + \beta_3\left(H_p - \bar{H}\right)F_p\right)\right)^2}{2\sigma_e^2}\right)$$

$$= \frac{1}{\sqrt{2\pi(4.73)}} \exp\left(-\frac{\left(W_p - \left(222.18 + 3.19\left(H_p - \bar{H}\right) - 82.27F_p - 1.09\left(H_p - \bar{H}\right)F_p\right)\right)^2}{2(4.73)}\right)$$

**Model for the Variance**

The plots were created using the following value for the predictors: $\bar{H} = 67.9$

Left plot: $H_p = 62; F_p = 1$

Right plot: $H_p = 76; F_p = 0$

# ML ESTIMATION OF GLMS: SAS PROC MIXED

# Maximum Likelihood Estimation for GLMs in SAS: PROC MIXED

- Maximum likelihood estimation of GLMs can be performed in SAS using PROC MIXED

- PROC MIXED will grow in value to you as time goes on: most multivariate analyses assuming conditional normality can be run with PROC MIXED:
  - Multilevel models
  - Repeated measures
  - Some factor analysis models

- The **MIXED** part of PROC MIXED refers to the type of model it can estimate: **General Linear Mixed Models**
  - Mixed models *extend* the GLM to be able to model dependency between observations (either within a person or within a group, or both)

# Likelihood Functions in PROC MIXED

- PROC MIXED uses a common (but very general) log-likelihood function based on the GLM: the conditional distribution of Y given **X**

$$f\left(Y_p \middle| X_p, Z_p\right) \sim N\left(\beta_0 + \beta_1 X_p + \beta_2 Z_p + \beta_3 X_p Z_p, \sigma_e^2\right)$$

  - Y is normally distributed conditional on the values of the predictors

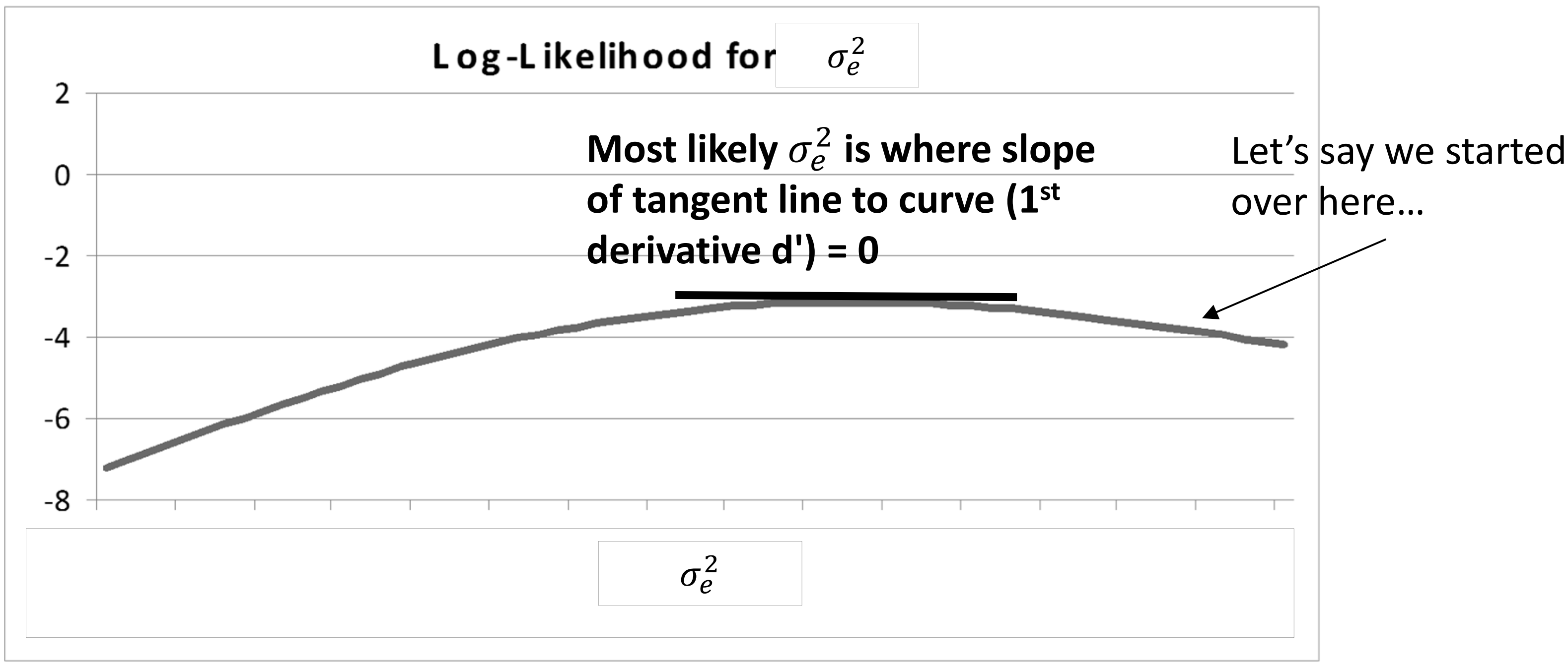
- The log likelihood for Y is then

$$\log L = \log L(\sigma_e^2 | x_1, \ldots, x_N) = -\frac{N}{2}\log(2\pi) - \frac{N}{2}\log(\sigma_e^2) - \sum_{p=1}^{N} \frac{\left(Y_p - \hat{Y}_p\right)^2}{2\sigma_e^2}$$

- Furthermore, there is a **closed form** (a set of equations) for the fixed effects (and thus $\hat{Y}_p$) for any possible value of $\sigma_e^2$
  - So…PROC MIXED seeks to find $\sigma_e^2$ at the maximum of the log likelihood function – and after that finds everything else from equations
  - Begins with a naïve guess…then uses Newton-Raphson to find maximum

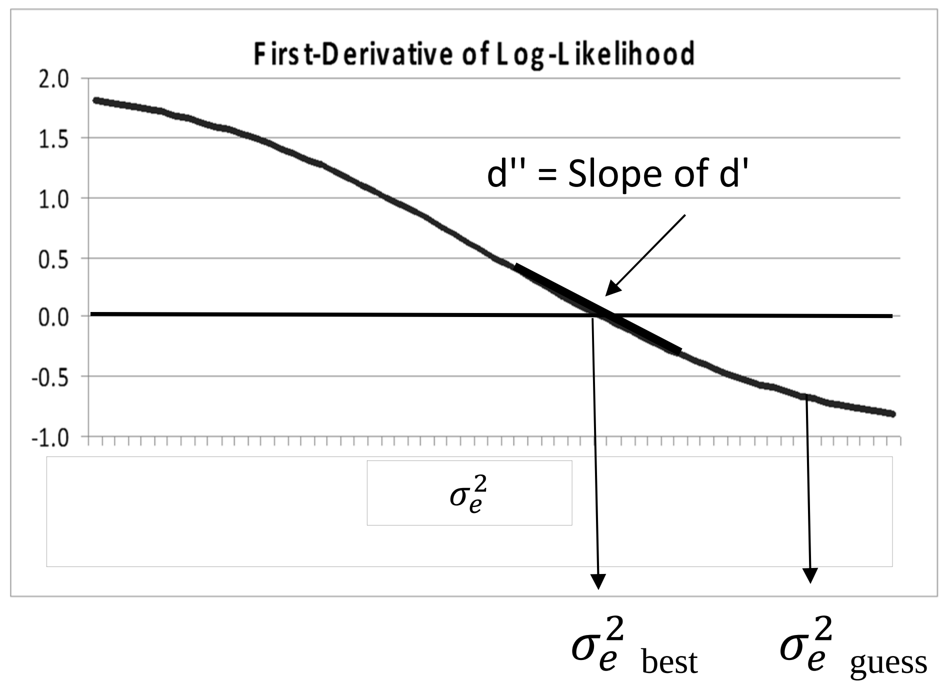
# $\sigma_e^2$ Estimation via Newton Raphson

- We could calculate the likelihood over wide range of $\sigma_e^2$ for each person and plot those log likelihood values to see where the peak is…
  - But we have lives to lead, so we can solve it mathematically instead by finding where the slope of the likelihood function (the 1st derivative, d') = 0 (its peak)
- Step 1: Start with a guess of $\sigma_e^2$, **calculate 1st derivative d'** of the log likelihood with respect to $\sigma_e^2$ at that point
  - Are we there (d' = 0) yet? Positive d' = too low, negative d' = too high



# $\sigma_e^2$ Estimation via Newton Raphson

- Step 2: **Calculate the 2nd derivative** (slope of slope, d'') at that point
  - Tells us **how far off we are**, and is used to figure out how much to adjust by
  - d'' will always be negative as approach top, but d' can be positive or negative
- Calculate new guess of $\sigma_e^2$ : $\sigma_{e\ new}^2 = \sigma_{e\ old}^2 - (d'/d'')$
  - If (d'/d'') < 0 → $\sigma_e^2$ increases
    If (d'/d'') > 0 → $\sigma_e^2$ decreases
    If (d'/d'') = 0 then you are done




- **2nd derivative d'' also tells you how *good* of a peak you have**
  - Need to know where your best $\sigma_e^2$ is (at d'=0), as well as how precise it is (from d'')
  - If the function is flat, d'' will be smallish
  - **Want large d'' because 1/SQRT(d'') = $\sigma_e^2$'s SE**

# USEFUL PROPERTIES OF MAXIMUM LIKELIHOOD ESTIMATES

# Likelihood Ratio (Deviance) Tests

- The likelihood value from MLEs can help to statistically test competing models assuming the models are nested

- Likelihood ratio tests take the ratio of the likelihood for two models and use it as a test statistic

- Using log-likelihoods, the ratio becomes a difference
  - The test is sometimes called a **deviance test**

$$D = \Delta - 2\log L = -2 \times (\log L_{H0} - \log L_{HA})$$

  - $D$ is tested against a Chi-Square distribution with degrees of freedom equal to the difference in number of parameters

# Deviance Test Example

- Imagine we wanted to test the null hypothesis that IQ did not predict performance:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

- The difference between the empty model and the conditional model is one parameter
  - Null model: one intercept $\beta_0$ and one residual variance $\sigma_e^2$ estimated = 2 parameters
  - Alternative model: one intercept $\beta_0$, one slope $\beta_1$, and one residual variance $\sigma_e^2$ estimated = 3 parameters

- Difference in parameters: 3-2 = 1 (will be degrees of freedom)

# Wald Tests (Usually 1 DF Tests in Software)

- For each parameter $\theta$, we can form the Wald statistic:

$$\omega = \frac{\hat{\theta}_{MLE} - \theta_0}{SE(\hat{\theta}_{MLE})}$$

  - (typically $\theta_0 = 0$)

- As N gets large (goes to infinity), the Wald statistic converges to a standard normal distribution $\omega \sim N(0,1)$
  - Gives us a hypothesis test of $H_0: \theta = 0$

- If we divide each parameter by its standard error, we can compute the two-tailed p-value from the standard normal distribution (Z)
  - Exception: bounded parameters can have issues (variances)

- We can further add that variances are estimated, switching this standard normal distribution to a t distribution (SAS does this for us)
  - Note: some don't like calling this a "true" Wald test

# Wald Test Example

- We could have used a Wald test to compare between the empty and conditional model, or:

$$H_0: \beta_1 = 0$$
$$H_1: \beta_1 \neq 0$$

- SAS provides this for us in the Solution for Fixed Effects:

| Solution for Fixed Effects | | | | | | |
|---|---|---|---|---|---|---|
| **Effect** | **Gender** | **Estimate** | **Standard Error** | **DF** | **t Value** | **Pr > \|t\|** |
| **Intercept** | | 222.18 | 0.8381 | 16 | 265.11 | <.0001 |
| **heightMC** | | 3.1897 | 0.1114 | 16 | 28.65 | <.0001 |
| **Gender** | F | -82.2719 | 1.2111 | 16 | -67.93 | <.0001 |
| **Gender** | M | 0 | . | . | . | . |
| **heightMC*Gender** | F | -1.0939 | 0.1678 | 16 | -6.52 | <.0001 |
| **heightMC*Gender** | M | 0 | . | . | . | . |

- Here, the interaction estimate has a t-test statistic value of -6.52 ($p < .0001$), meaning we would reject our null hypothesis

- Typically, Wald tests are used for one additional parameter

# Information Criteria

- Information criteria are statistics that help determine the relative fit of a model for non-nested models
  - Comparison is fit-versus-parsimony
- PROC MIXED reports a set of criteria (from conditional model)

| Information Criteria | | | | | | |
|---|---|---|---|---|---|---|
| Neg2LogLike | Parms | AIC | AICC | HQIC | BIC | CAIC |
| 86.5 | 1 | 88.5 | 88.8 | 88.6 | 89.3 | 90.3 |

  - Each uses -2*log-likelihood as a base
    - Choice of statistic is **very** arbitrary and depends on field

- Best model is one with *smallest* value

- Note: don't use information criteria for nested models
  - LRT/Deviance tests are more powerful

Adding to the fun…

# REML ESTIMATION IN MULTILEVEL MODELS

# What about ML vs. REML?

- REML estimates of random effects variances and covariances are unbiased because they account for the uncertainty that results from simultaneously also estimating fixed effects (whereas ML estimates do not, so they are too small)

- What does this mean? Remember "population" vs. "sample" formulas for computing variance?

| **"Population"** | **"Sample"** |
|---|---|
| $\sigma_e^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N}$ | $\sigma_e^2 = \frac{\sum_{i=1}^{N}(y_i - \bar{y})^2}{N-1}$ |

  - N-1 is used because the mean had to be estimated from the data

- Same idea: ML estimates of random effects variances will be downwardly biased by a factor of (N – k) / N, where N = # persons and k = #fixed effects…
  it just looks way more complicated

# What about ML vs. REML?

$$\text{ML:} \quad LL = \left[-\frac{T-0}{2}\log(2\pi)\right] + \left[-\frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{V}_i|\right] + \left[-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\gamma})^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\gamma})\right]$$

$$\text{REML:} \; LL = \left[-\frac{T-k}{2}\log(2\pi)\right] + \left[-\frac{1}{2}\sum_{i=1}^{N}\log|\mathbf{V}_i|\right] + \left[-\frac{1}{2}\sum_{i=1}^{N}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\gamma})^T \mathbf{V}_i^{-1}(\mathbf{Y}_i - \mathbf{X}_i\boldsymbol{\gamma})\right] + \left[-\frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right|\right]$$

$$\text{where:} \quad \left[-\frac{1}{2}\log\left|\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right|\right] = \left[\frac{1}{2}\log\left|\left(\sum_{i=1}^{N}\mathbf{X}_i^T\mathbf{V}_i^{-1}\mathbf{X}_i\right)^{-1}\right|\right] = \underbrace{\left[\frac{1}{2}\log|\mathrm{Cov}(\boldsymbol{\gamma})|\right]}$$

- Extra part in REML is the sampling variance of the fixed effects… it is added back in to account for uncertainty in estimating fixed effects
- REML maximizes the likelihood of the residuals specifically, so models with different fixed effects are not on the same scale and are not comparable
  - This is why you can't do -2ΔLL tests in REML when the models to be compared have different fixed effects as the model residuals are defined differently

# Wrapping Up

- This section was an introduction to mathematical statistics as a way to understand the implications statistical models make about data

- Although many of these topics do not seem directly relevant, they help provide insights that untrained analysts may not easily attain
  - They also help you to understand when and when not to use a model!