Introduction to this Course and to Maximum Likelihood Estimation of Generalized Linear Models

- Topics:
  - Course overview
  - Quantitative methods: A Lego-inspired world view
  - Review of general linear models, my way
  - A little bit about:
    - Maximum likelihood
    - Link functions
    - Outcome variable types and conditional distributions
    - Multivariate modeling and software
    - Why this class will help you in the future

What To Expect This Semester…

- You are here to expand your knowledge of **quantitative methods** = quantitative data + application of statistics to answer questions
  - Models are the lens though which we view research
  - New models → new questions → new answers

- This will NOT require anxiety-provoking behaviors like:
  - Calculating things by hand—computers are always better, and more advanced statistical methods cannot be implemented by hand anyway
  - Deriving formulas or results—it's ok to trust the people who specialize in these areas to have gotten it right and use their work (for now, at least)
  - Memorizing formulas—it's ok to trust the computer programmers who have implemented various statistical techniques (for now, at least)

- It WILL require learning and implementing **new language and decision guidelines** for matching data, questions, and models

How to Expand Your Knowledge of Quantitative Methods

- I will NOT:
  - Use infrequent high-stakes testing to assess your level of learning
  - Ask you to complete practice problems that have nothing to do with the research process
  - Present statistics as a series of unrelated ideas and formulae

- I WILL:
  - Use **formative assessments** to help you figure out what you need to review (7 planned; 14 points for **completing them at all**)
  - Require online **homework assignments** that give you real-world practice (6 planned; 86 points for **completing them accurately**)
  - Present statistics by linking data, questions, and models explicitly

Our Responsibilities

- My job:
  - Provide custom lecture materials and examples that are accurate, comprehensive, and with the necessary scaffolding for your future use
  - Answer questions via email, in individual meetings, or in group-based office hours—you are ALL invited to attend to work on homework during office hours and get immediate assistance if you want it

- Your job:
  - **Ask questions**—preferably in class, but any time is better than none
  - **Review** the class material **frequently**, focusing on mastering the vocabulary (words and symbols), logic, and procedural skills
  - **Practice** using the software to implement the techniques you are learning **on data you care about**—this will help you so much more!
  - Read the texts *if you feel they are helpful* (they are mainly for reference)
  - Don't wait until the last minute to start homework, and don't be afraid to **ask for help** if you get stuck on one thing for more than 15 minutes

Class-Sponsored Statistical Software

- To help address the needs of different programs, I will show examples using **SAS, STATA, and likely M*plus*** (Why? Stay tuned…)
  - Why not SPSS? Because it doesn't have everything we need and it doesn't leave as much room to grow into advanced models
  - Caveat: I am a heavy-duty SAS user who picked up enough STATA to teach multilevel modeling workshops using it
    - So if you have STATA tips, please share them with me!
  - Btw, I can also help you in a little bit of R (and so can the Agresti book)

- Things to consider when choosing which one to focus on:
  - More programs = more entries in your "technical skills" part of CV
  - Although SAS and STATA are available through the UIowa Virtual Desktop, **only SAS is available from off campus, too**
  - What program will be used in your quant classes to follow?
  - What do the other members of your research lab use?

SAS vs. STATA: My Opinion

Activity	Winner	Commentary
Working with raw files or multiple datasets	SAS, hands down	As of STATA 15, only one dataset can be open at once—problematic for messy data management
Within-dataset manipulations	Tie, but STATA for some tasks	STATA wins for group-centering, stacking, and unstacking (used for multilevel models)
Data analysis	Tie, but SAS for some tasks	I've had estimation problems in STATA for certain advanced model variants (within multilevel models)
Post-estimation (i.e., predicted outcomes or simple slopes)	STATA, hands down	STATA has simple yet powerful options for doing these tasks in bulk that SAS doesn't have
Automating data tasks (i.e., loops)	Tie	Both programs have ways to do this, but I only know how in SAS…

This Semester's Topics

- **Generalized linear models for univariate outcomes:**
  - Conditionally normal outcomes (review of General Linear Models)
  - Binary and Categorical outcomes
  - Count and "If-and-How-Much" outcomes
  - Non-normal continuous outcomes (and maybe Quantile Regression)

- **Models for multivariate and repeated measures outcomes**
  - Conditionally normal outcomes
  - Other non-normal outcomes
- **Path analysis for "truly" multivariate and mediation analysis**
  - Conditionally normal outcomes
  - Other non-normal outcomes

- But first, the bigger picture and some background…

Quant Methods: A Lego-Based Approach

My goal today:

a) describe these **4 Legos**

b) use them to provide the "big picture" of this course

Big Picture Idea:
If you understand the elemental building blocks of statistical models, then you can build **anything**!

The Origins of these Legos

- Problem: The **giant canyon** between two types of classes
  → To cross it, students need **2 kinds** of training
  - Become conversant in **traditional** methods (and the terms that go with them) still commonly used in many research areas
  - Recognize the **building blocks** of modern analytic techniques (current and future) to build a pathway to fluency with them
    - Recognizing the building blocks of traditional methods helps, too

- Solution: Build a **bridge course** that crosses this canyon
  - In specific: PSQF 7375, Applied Generalized Linear Models
  - In general: A Lego-based **philosophy** for learning quantitative methods developed in cahoots with Jonathan Templin

The 4 Lego Building Blocks

The Legos we will learn in this course…

1. **Linear models** (for **answering questions** of prediction)
2. **Estimation** (for iterative ways of **finding the answers**)
3. **Link functions** (for predicting **any type of outcome**)

…will better prepare you for other courses also with:

4. (a) **Random effects** / (b) **Latent variables**

    (a) for modeling multivariate **"correlation/dependency"** (using multilevel or mixed-effects models)

    (b) for modeling relations of **"unobserved constructs"** (using factor analysis, item response models, or SEM)

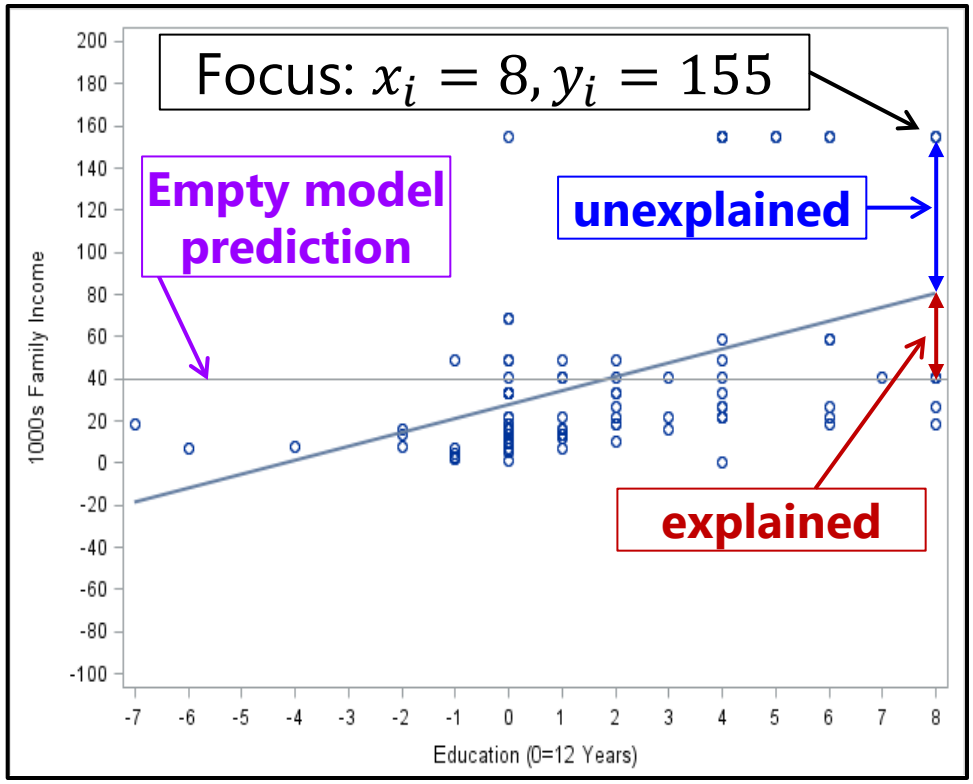
1. Linear Models Run the World

- **Linear models are the mechanism** by which the vast majority of all research questions will be answered
  - *Is there an effect? Is this effect the same for everyone? Is the effect still there after considering something else?*

- A linear-models world view entails starting with the most **traditional models**, but from a **different perspective**
  - More intuitive: linear regression models
    - *Because the focus is on the fixed effects in the model equation*
  - Less intuitive: analysis of variance in group-based designs
    - *Because the focus is on cell and marginal mean differences (which are indirectly provided by the model fixed effects)*
  - Each of these is one flavor of the **General Linear Model**…

A One-Slope GLM Example

The $\boldsymbol{\beta}$ estimates result from the goal of minimizing the sum of squared residuals across the sample—this is "**ordinary least squares estimation**"—let's see what happens for one person:



$\boldsymbol{\beta_0}$ = **intercept,** $\boldsymbol{\beta_1}$ = **slope**

Empty Model for y_i = Income:

$$\boldsymbol{y_i = \beta_0 + e_i}$$

$$\boldsymbol{\hat{y}_{Focus} = 40}$$

$$\boldsymbol{y_{Focus} = 40 + 115}$$

Variance: $\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-1} = 1980$

Add Education (centered so that 12 years = 0) as Predictor:

$$\boldsymbol{y_i = \beta_0 + \beta_1(Educ_i - 12) + e_i}$$

$$\boldsymbol{\hat{y}_{Focus} = 27 + 6.6(8) = 80}$$

$$\boldsymbol{y_{Focus} = 80 + 75}$$

Variance: $\boldsymbol{\sigma_e^2} = \frac{\sum_{i=1}^{N}(y_i - \hat{y}_i)^2}{N-2} = 1594$

General Linear Models, More Generally

- A **General Linear Model (GLM*)** for outcome $\boldsymbol{y_i}$ looks like this:
  - actual $\boldsymbol{y_i = \beta_0 + \beta_1(x1_i) + \beta_2(x2_i) + \cdots \beta_p(xp_i) + e_i}$
  - predicted $\boldsymbol{\hat{y}_i = \beta_0 + \beta_1(x1_i) + \beta_2(x2_i) + \cdots \beta_p(xp_i)}$

  - The "i" subscript denotes **variables** (that are individual-specific)
  - The $\boldsymbol{\beta}$ ("beta") terms are the model **fixed effects** → **constants** whose subscripts range from 0 up to p as the last fixed effect):
    - $\boldsymbol{\beta_0}$ = **intercept** = expected $\boldsymbol{y_i}$ when all x_i predictors are 0
    - $\boldsymbol{\beta_1}$ = **slope of** $\boldsymbol{x1_i}$ = difference in $\boldsymbol{y_i}$ per one-unit difference in $\boldsymbol{x1_i}$
    - $\boldsymbol{\beta_2}$ = **slope of** $\boldsymbol{x2_i}$ = difference in $\boldsymbol{y_i}$ per one-unit difference in $\boldsymbol{x2_i}$
    - **…**
    - $\boldsymbol{\beta_p}$ = **slope of** $\boldsymbol{xp_i}$ = difference in $\boldsymbol{y_i}$ per one-unit difference in $\boldsymbol{xp_i}$

***** GLM may also stand for Generalized Linear Models, which includes General as one type

How Many Fixed Effects Per Predictor?

- The **role of each predictor x_i** in creating a custom expected outcome y_i is described through one or more fixed slopes
  - **One slope** is sufficient to capture the mean difference between two categories for a **binary x_i** or to capture a **linear effect** of a quantitative x_i (or exponential for log x_i or logistic for logit x_i)
  - **More than one slope** may be needed to capture other nonlinear effects of a quantitative x_i (e.g., **quadratic** or **piecewise** trends)
  - **$C - 1$ slopes** are needed to capture the mean differences in the outcome across a **categorical predictor** with **C categories**

- "**Linear**" in GLM refers to "slope*variable + slope*variable" format
  - This means the x_i predictors can also be nonlinear terms (like x_i^2 to create a curve for x_i), which is then called "**nonlinear in the variables**"
  - The alternative, "**nonlinear in the parameters**" would have a nonlinear form, e.g., this exponential model: $\hat{y}_i = \beta_0 + \beta_1[exp(\beta_2(x1_i))]$

Testing Significance of Fixed Effects

- Any single-df **fixed effect** has 4-5 relevant pieces of output:
  - **Estimate** = best guess for the fixed effect based on our data
  - **Standard Error** = index of the precision of fixed effect estimate (i.e., quality of the "most likely" estimate)
  - ***t*-value or *z*-value** = Estimate / Standard Error
  - ***p*-value** = probability that fixed effect estimate is ≠ 0
  - **95% Confidence Interval** = Estimate ± 1.96*SE = range in which true (population) value of estimate is expected to fall 95% of the time

- Compare test statistic (*t* or *z*) to critical value at chosen level of significance (known as alpha): this is a "**univariate Wald test**"
- Whether the *p*-value is based on *t* or *z* varies by program…

Testing Significance of Fixed Effects

Fixed effects can be tested via **Wald** tests: the ratio of its estimate/SE forms a statistic we compare to a distribution

	Denominator DF is infinite (Proper Wald test)	Denominator DF is estimated instead ("Modified" Wald test)
Numerator DF = 1 *(test one fixed effect)* is **Univariate Wald Test**	use ***z*** distribution (Mplus, STATA)	use ***t*** distribution (SAS, SPSS)
Numerator DF > 1 *(test 2+ fixed effects)* is **Multivariate Wald Test**	use χ^2 distribution (Mplus, STATA)	use ***F*** distribution (SAS, SPSS)
Denominator DF options (in Stata use "small", not a default)	not applicable, so DDF is not given	SAS, STATA 14: BW, KR SAS, STATA 14, SPSS: Satterthwaite

Significance of Each Fixed Slope

- Standard Error (SE) for fixed effect estimate β_X in a one-predictor model (remember, SE is like the SD of the estimated parameter):

$$SE_{\beta_X} = \sqrt{\frac{\text{residual variance of Y}}{\text{Var(X)} * (N-k)}}$$

N = sample size
k = number of fixed effects

- When more than one predictor is included, SE turns into:

$$SE_{\beta_X} = \sqrt{\frac{\text{residual variance of Y}}{\text{Var(X)} * \mathbf{(1-R_X^2)} * (N-k)}}$$

R_X^2 = X variance accounted for by other predictors, so **$1-R_X^2$ = unique X variance**

- So all things being equal, SE is smaller when:
  - More of the outcome variance has been reduced (better predictive model)
    - This means fixed effects can become significant later if R^2 is higher then
  - The predictor has less covariance with other predictors
    - Best case scenario: X is uncorrelated with all other predictors

- If SE is smaller → *t*-value or *z*-value is bigger→ *p*-value is smaller

Multivariate Wald Tests of Fixed Effects

- General test for significance of **multiple fixed effects** at once (can be requested via CONTRAST in SAS, or TEST in STATA and SPSS) —you have likely already seen these special cases…

- Whether the set of fixed slopes for x_i significantly explains y_i variance (i.e., if $R^2 > 0$) is tested via "**Multivariate Wald Test**"
  - $F(DF_{num}, DF_{den}) = \frac{SS_{model}/(k-1)}{SS_{residual}/(N-k)} = \frac{(N-k)R^2}{(k-1)(1-R^2)} = \frac{known}{unknown}$
  - $\boldsymbol{F}$**-test** evaluates model R^2 *per DF spent to get it and DF leftover*
  - $\boldsymbol{R^2} = \frac{SS_{total}-SS_{residual}}{SS_{total}}$ = square of r between predicted $\boldsymbol{\hat{y}_i}$ and $\boldsymbol{y_i}$

- "Omnibus" F-test for the slopes of the main effect of a variable with $C > 2$ categories (or for its interaction with other predictors)
- Model R^2 change F-test in hierarchical regression (for grouping sets of predictors together and testing their joint contribution)

A Taxonomy of Fixed Effect Interpretations

- After significance testing comes interpretation. Fixed effects will be either:
  - an **intercept** that provides an expected (conditional) $\boldsymbol{y_i}$ outcome,
  - or **a slope** for the expected difference in $\boldsymbol{y_i}$ per unit difference in $\boldsymbol{x_i}$

- **All slopes** can be described as falling within one of three categories: ***bivariate marginal***, ***unique marginal***, or ***unique conditional***
  - In models with only **one fixed slope**, that slope's main effect is ***bivariate marginal*** (is uncontrolled; applies across all persons)
  - In models with **more than one fixed slope**, each slope's main effect is ***unique*** (it controls for the overlap in contribution with each other slope)
    - If a predictor is not part of an interaction term, its ***unique effect is marginal*** (it controls for the other slopes, but still applies across all persons)
    - If a predictor is part of one or more interaction terms, its ***unique effect is conditional*,** which means it is **specific to each interacting predictor = 0**
      - **Unique conditional** effects are also called **"simple main effects"**

Practice Labeling Fixed Effects

Model: $\boldsymbol{y_i = \beta_0 + \beta_1(F_i) + \beta_2(G_i) + \beta_3(H_i) + \beta_4(G_i)(H_i) + e_i}$
Choices: bivariate marginal, unique marginal, unique conditional

- Label for effect of F =
  - Equation? Effect of F =
- Label for effect of G =
  - Equation? Effect of G =
- Label for effect of H =
  - Equation? Effect of H =
- For practice with these concepts and accompanying syntax, see recommendations posted with lecture 0 from my other classes…

Flavors of General Linear Models

- Unlike any other family of statistical models, **the same General Linear Model is called different names** (often taught in different classes) based on **what kind of x_i predictor variables** are included:
  - One quantitative predictor? "Simple (linear) regression"
  - 2+ quantitative predictors? "Multiple (linear) regression"
  - One categorical predictor with two groups? "Independent t-test"
  - One categorical predictor with 3+ groups? "One-way ANOVA"
  - 2+ categorical predictors (with interactions between them)? "Two-way (or more-way) ANOVA"
  - 2+ categorical predictors (with interactions between them) and 1+ quantitative predictors (without interactions with the categorical predictors)? "Two (or more)-way ANCOVA"
  - Whatever combination is necessary? "Multiple regression"
- These distinctions only serve to confuse people and obfuscate what is **just one model**, the General Linear Model… **here is why**:

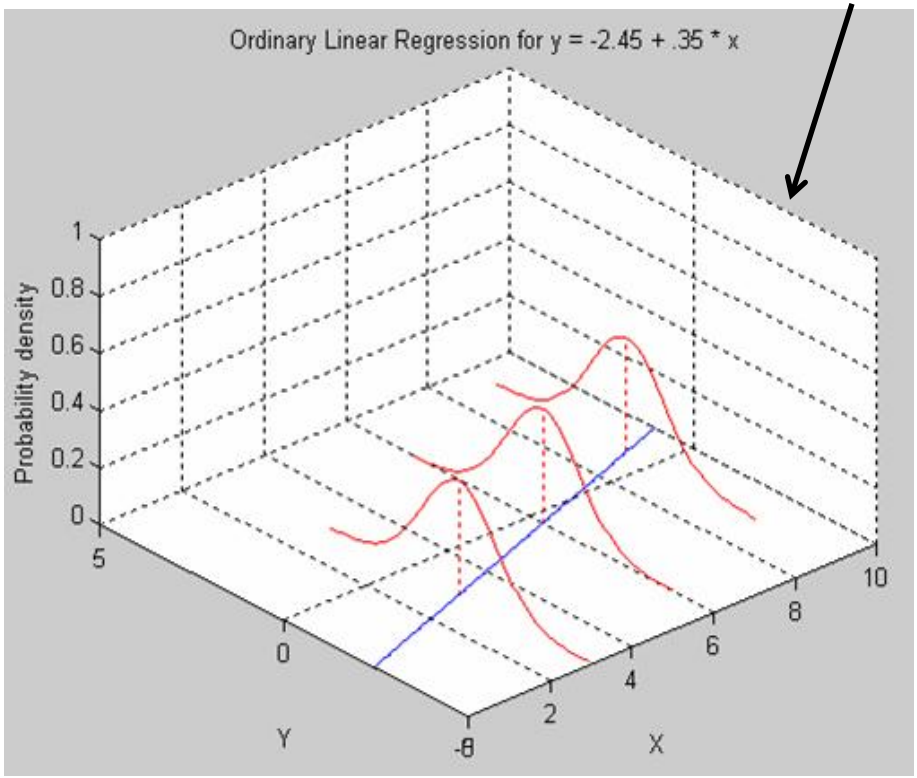
General Linear Model Residuals

- GLM for actual $\boldsymbol{y_i = \beta_0 + \beta_1(x1_i) + \beta_2(x2_i) + \cdots \beta_p(xp_i) + e_i}$
- GLM for predicted $\boldsymbol{\hat{y}_i = \beta_0 + \beta_1(x1_i) + \beta_2(x2_i) + \cdots \beta_p(xp_i)}$

- No matter what kind of predictors (and whether or not their interactions) are included, the term "**General**" in GLM refers to the use of a **conditional normal distribution** for the $\boldsymbol{e_i}$ residuals, in which $\boldsymbol{e_i}$ = actual $\boldsymbol{y_i}$ − predicted $\boldsymbol{\hat{y}_i}$
  - This "general" idea is written formally like this $\boldsymbol{y_i \sim \mathbf{N}(\hat{y}_i, \sigma_e^2)}$: $\boldsymbol{y_i}$ is *Normally* distributed with $Mean = \boldsymbol{\hat{y}_i}$ and $Variance = \boldsymbol{\sigma_e^2}$
  - In addition, in the GLM, the $\boldsymbol{e_i}$ **residuals are assumed independent**, (although in many types of research designs this cannot be true)
  - Further, everyone with the same combination of x_i predictor values would have the same $\boldsymbol{\hat{y}_i}$, and the **model predicts equally well** for everyone (because there is **only one residual variance**, $\boldsymbol{\sigma_e^2}$)

General Linear Model Residual Variance

- The GLM assumes equal (constant) residual variability across all predictor values: "**homoscedasticity**" = "**homogeneity of variance**"



Otherwise, "**heteroscedasticity** " = "**heterogeneity of variance**" → model predicts differentially well across x_i (SEs will need adjusted)

"Not good" → $\boldsymbol{\sigma_e^2}$ increases as the x_i predictor increases (→ fan)

Solution: Add fixed effects that allow the variance to differ (this **leaves GLM**)

Left image borrowed from: http://www.omidrouhani.com/research/logisticregression/html/logisticregression.htm
Right image borrowed from: https://ajh1143.github.io/HomVar/

How the Lego Blocks Fit Together

1. **Linear models** answer research questions, and are the first building block of every more complex analysis
   - *Is there an effect? Is this effect the same for everyone? Is the effect still there after considering something else?*

What other blocks you will need is determined by:

3. How your outcome is measured → **link functions**
4. Your dimensions of sampling → **random/latent effects**

- How can we add these Legos? → **2. new estimation**
  - **Least squares** is taught first, but is greatly limited in practice
  - **Maximum likelihood** picks up where least squares leaves off
  - **Bayesian** picks up where maximum likelihood gives up

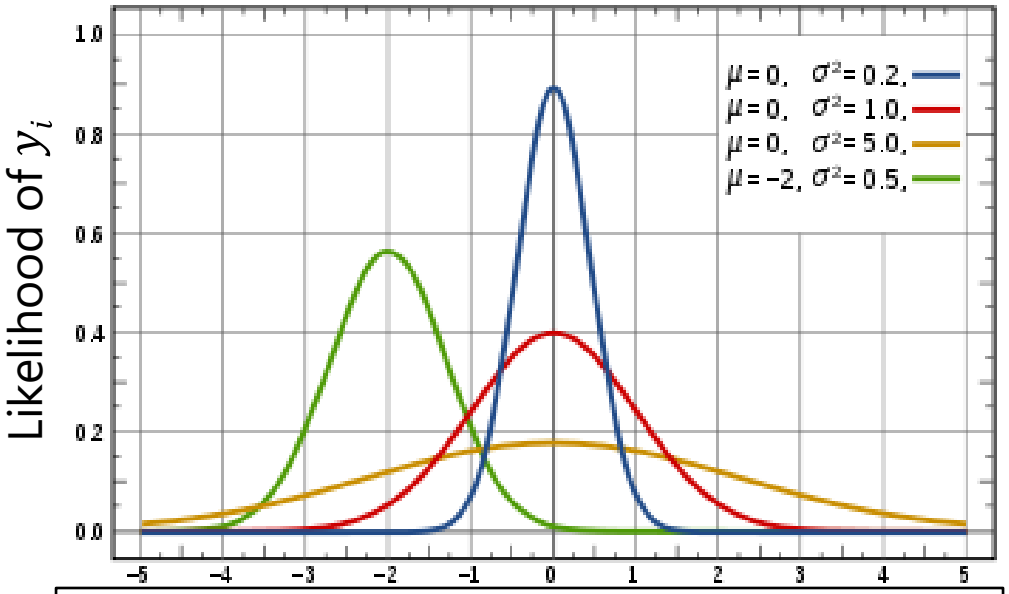
2. Estimation via Maximum Likelihood

- Ordinary Least Squares (OLS) can find answers in **some** kinds of data
  - "Best" fixed effects are those that minimize the sum of squared errors
  - How? Calculate sums of squares → mean squares → F-ratios…

- The good news: **Maximum likelihood (ML) can find the answers** with more flexibility in **many more kinds of data**
  - Non-normal, multivariate, clustered, or incomplete data… in fact, an ML variant called *residual ML* (or *REML*) simplifies to least squares
  - OLS calculations are computational shortcuts to REML (see Enders ch. 3)

- **The even better news**: If you understand **this**, then you understand the basics of ML
  - Can still work with some calculations for pedagogical purposes, though, like this…

Univariate Normal **P**robability **D**istribution **F**unction



Univariate Normal PDF:

$$f(y_i) = \frac{1}{\sqrt{2\pi\sigma_e^2}} * \exp\left[-\frac{1}{2} * \frac{\left(y_i - \hat{y}_i\right)^2}{\sigma_e^2} \right]$$

Sum over persons of log of $f(y_i)$= Model Log-Likelihood → Model Fit

- This PDF tells us how **likely** (i.e., **tall**) any value of $\boldsymbol{y_i}$ is given two things:
  - Conditional mean $\boldsymbol{\hat{y}_i}$
  - Residual variance $\boldsymbol{\sigma_e^2}$
- We can see this work using the NORMDIST function in excel!
  - Easiest for **empty** model:

$$\boldsymbol{y_i = \beta_0 + e_i}$$

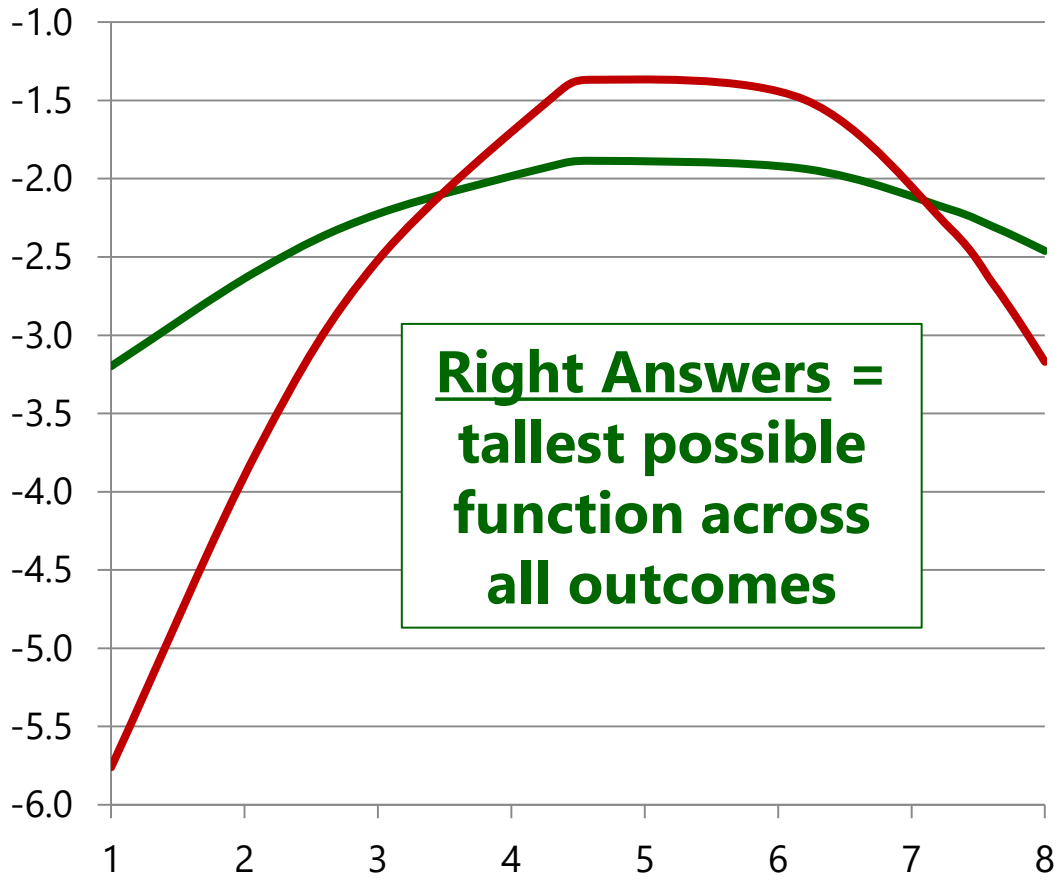
- We can check our math via software using ML!

Image borrowed from: https://en.wikipedia.org/wiki/Normal_distribution

ML via Excel NORMDIST

Key idea: Normal Distribution formula → data height

Mean	**5.19**	**5.24**
Variance	**6.56**	**2.00**
	Right	**Wrong**
Outcome	**Log(Height)**	**Log(Height)**
1.0	**-3.20**	**-5.76**
2.1	**-2.59**	**-3.73**
3.0	**-2.22**	**-2.52**
4.3	**-1.92**	**-1.49**
4.6	**-1.89**	**-1.37**
6.2	**-1.94**	**-1.50**
7.3	**-2.20**	**-2.33**
7.6	**-2.30**	**-2.66**
7.8	**-2.38**	**-2.90**
8.0	**-2.46**	**-3.17**
SUM = Model LL = taller is better		
	-23.09	**-27.42**



What's so great about normal?

- Why must we assume "**normality**, **independence**, and **constant variance**" of residuals in General Linear Model? Because those are **required by the formula it uses** to calculate each outcome's height!
  - The normal distribution only has one variance that is shared over people
  - Summing the log-likelihood over persons implies independent values

- **The magic of ML**: if your residuals aren't normally distributed, then you can just **pick a different formula for height**, such as one that:
  - Has a better-suited probability distribution for non-normal outcomes
  - Includes a linear model for heterogeneity of variance across people
  - And/or uses a multivariate version instead for dependent outcomes

3. Then, link functions to the rescue!

- Linear models + ML + link functions = generalized models
- But first, *what other types of outcomes (and distributions) are there???*

Other Types of Outcome Variables

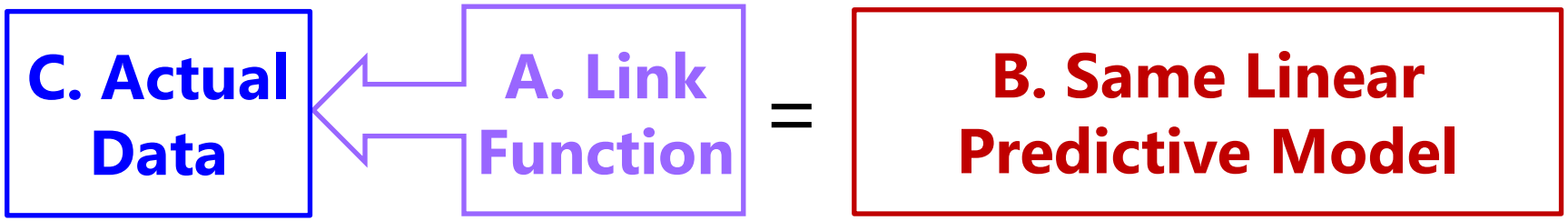
* Note: this is related to traditional levels of measurement, but I am approaching it from more of a "how-to-model them" perspective

- First, **categorical variables: *where the numbers are labels***
  - Binary (dichotomous) = 2 choices (typically coded as 0 or 1)
    - e.g., dead or alive; pregnant or not
  - Nominal = 3+ unordered choices
    - e.g., favorite type of pet
  - Ordinal = 3+ choices with some natural (undeniable) order, but the distances between the values used don't mean anything
    - e.g., 1 = strongly disagree, 2 = disagree, 3 = agree, 4 = strongly agree
    - Equally ordinal (and equally acceptable) values: 1, 20, 300, 4000
  - Synonyms for a "**categorical**" variable: discrete variable, qualitative variable, grouping variable, factor variable, CLASS variable (in SAS)

Other Types of Outcome Variables

- Next, **quantitative variables** where the **numbers are really numbers** (interval measurement → equal distances between all sets of values), but that have one or more natural boundaries
  - Binomial = number of occurrences out of known possible
    - e.g., # correct on a test, which is bounded by 0 and total possible
    - Correcting for different totals possible by computing proportion correct (or rate of occurrence) is still binomial (just bounded by 0 and 1 instead)
    - Scale sums with observed boundaries may also look like a binomial
  - Count = number of occurrences out of unknown possible
    - # of cigarettes smoked each day (minimum = 0, but maximum could be any positive number)
  - Count variables have special cases involving zero values:
    - No zeros possible? → zero-*truncated* count
    - More zeros than expected? → zero-*inflated* count ("if and how much")

3 Parts of Generalized Linear Models



A. **Link Function**: Transformation of conditional mean to keep predicted outcomes within the bounds of the outcome

B. **Same Linear Model**: How the model linearly predicts the *link-transformed* conditional mean of the outcome

C. **Conditional Distribution**: How the outcome residuals could be distributed given the possible values of the outcome

Generalized linear models work for many kinds of outcomes…

Quick Example for Binary Outcomes

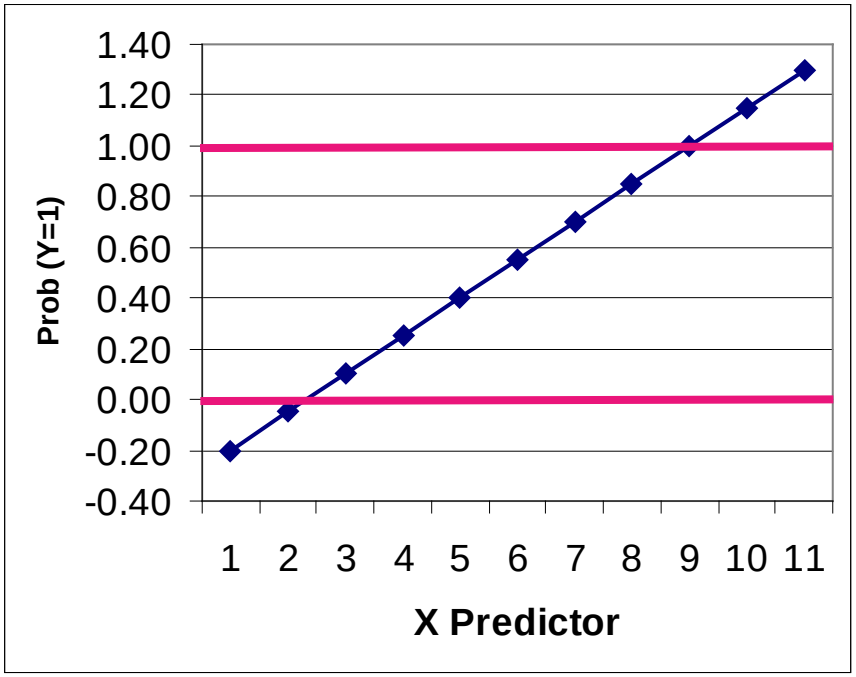
We need to go from this ***unbounded* linear model** for predicting probability…

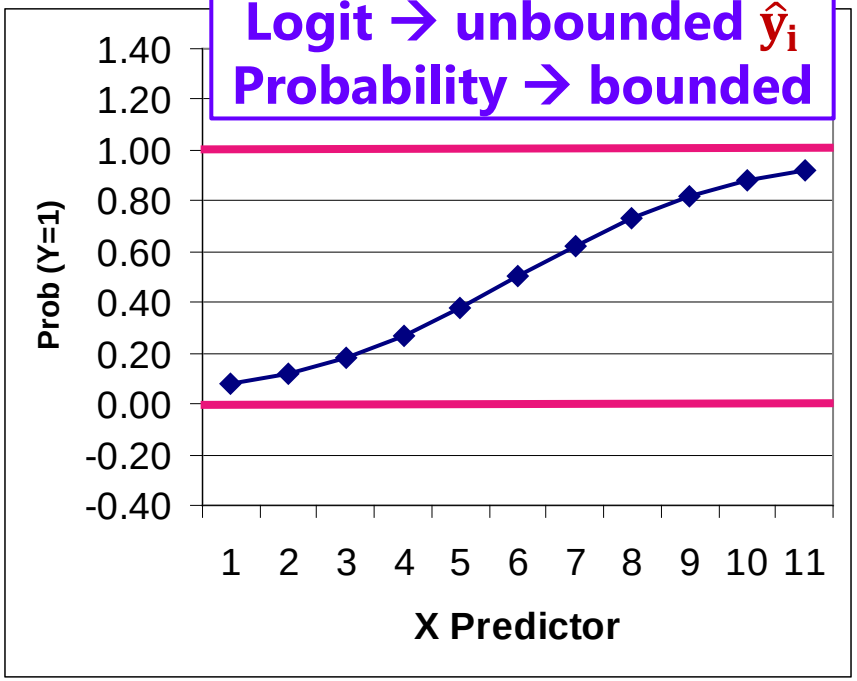
$$p(y_i = 1) = \beta_0 + \beta_1(x_i)$$

To this…

Logit Link

$$\text{Log}\left(\frac{p(y_i = 1)}{p(y_i = 0)}\right) = \beta_0 + \beta_1(x_i)$$



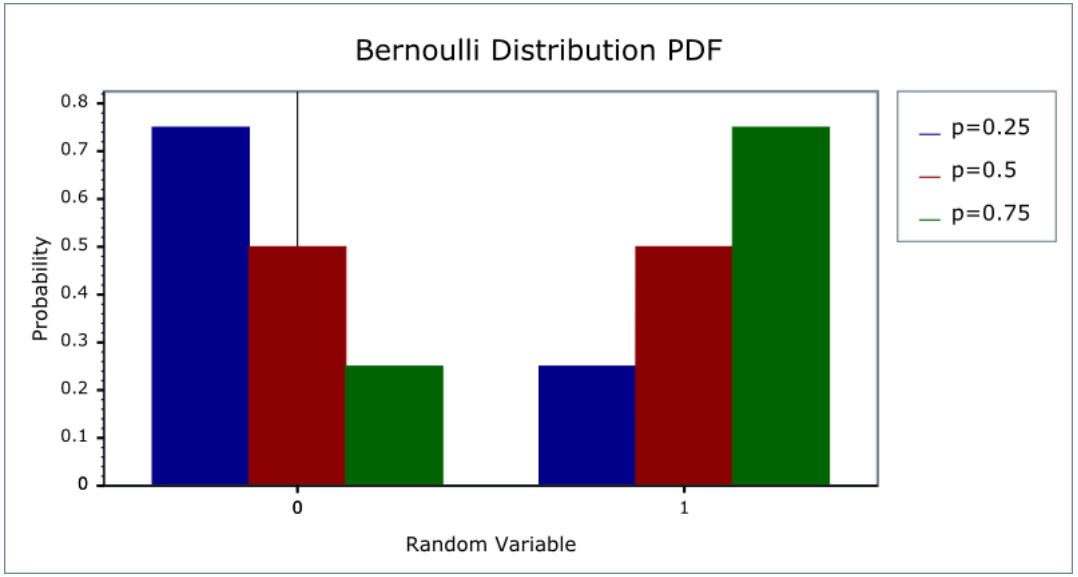


3. Link Functions help fix “Messy” Data

- Many kinds of **non-normal outcomes** can be analyzed with generalized models through the **magic of ML**

- **Two parts: Link function + other conditional distribution**
  - **Binary** → **Logit** + **Bernoulli**
  - **Ordinal** or **Nominal** → **Logit** + **Multinomial**
  - **% Correct** → **Logit** + **Binomial** (that have floor or ceiling effects!)
  - **Bimodal** → **Logit** + **Beta**
  - **Counts** → **Log** + **Poisson**
  - **Skewed Counts** → **Log** + **Negative Binomial**
  - **Skewed Continuous** → **Log** + **Log-Normal/Gamma**
  - **Zero-Inflated** (if and how much) → **Logit/Log** + **Bernoulli/other**

Bernoulli Distribution: Binary Variables



Bernoulli PDF:

$$f(y_i) = (p_i)^{y_i} (1 - p_i)^{1 - y_i}$$

= $p(1)$ if 1, → p
$p(0)$ if 0 → q

The Bernoulli distribution has only one parameter, called p, which is the mean: the proportion of 1 values (and $1 - p = q$).

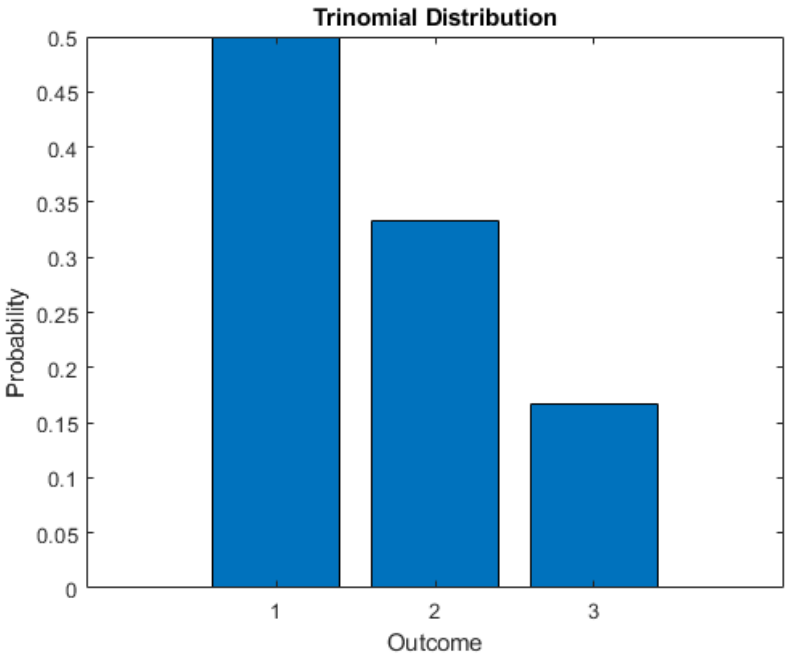
The mean determines variance = $p * q$ (and skewness = $\frac{1-2p}{\sqrt{p*q}}$)

Mean and Variance of a Binary Variable

Mean (p)	.0	.1	.2	.3	.4	.5	.6	.7	.8	.9	1.0
Variance	.0	.09	.16	.21	.24	.25	.24	.21	.16	.09	.0

Image borrowed from: https://www.boost.org/doc/libs/1_70_0/libs/math/doc/html/math_toolkit/dist_ref/dists/bernoulli_dist.html

Multinomial Categorical Distribution: Nominal or Ordinal Variables



- For example, $C = 3$ possible responses of $c = 1,2,3$, an observed $y_i = c$, and indicators I if $c = y_i$

$$f(\mathrm{y_i} = \mathrm{c}) = p_{\mathrm{i1}}^{\mathrm{I[y_i=1]}} p_{\mathrm{i2}}^{\mathrm{I[y_i=2]}} p_{\mathrm{i3}}^{\mathrm{I[y_i=3]}}$$

$= p_1$(1) if 1,
p_2(1) if 2,
$1 - (p_1 + p_2)$ if 3

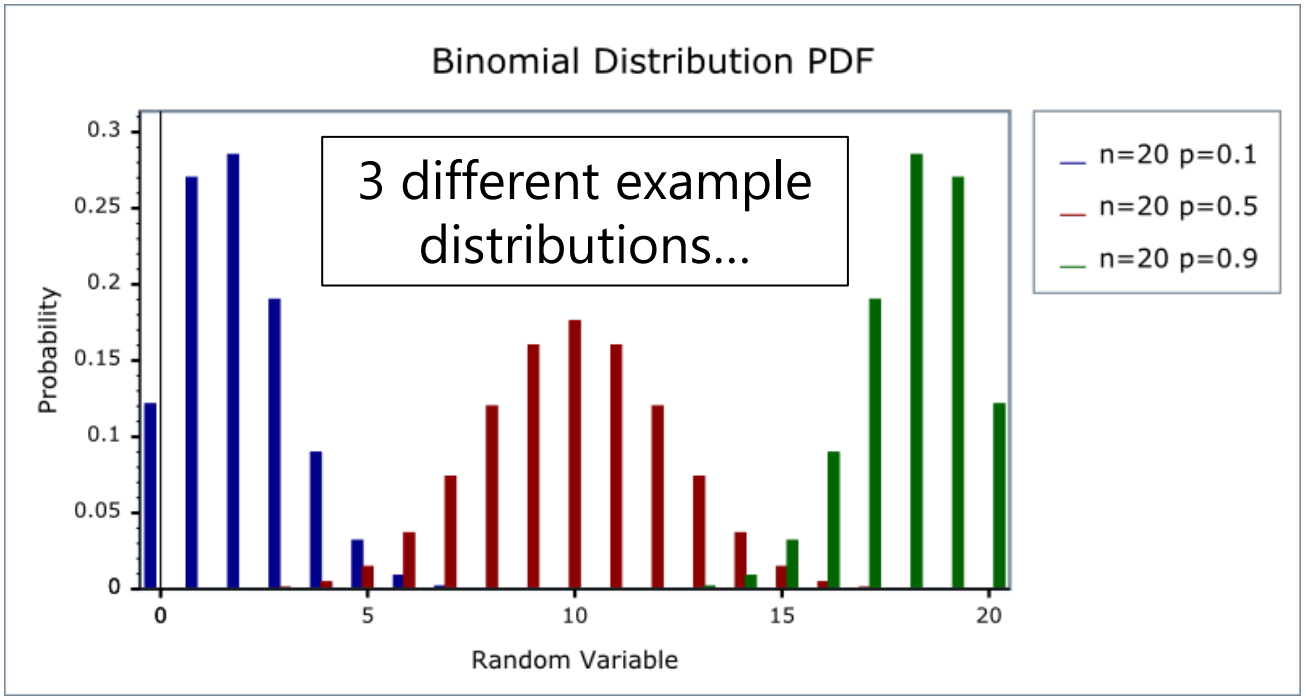
The multinomial distribution has $C - 1$ p mean parameters, called p_c, which create the proportion in each category (so variance is not a separate thing)

Image borrowed from: https://www.mathworks.com/help/stats/work-with-multinomial-probability-distribution-objects.html

Binomial Distribution: Proportions

- The discrete **binomial** distribution can be used to predict $\boldsymbol{c}$ correct responses given $\boldsymbol{n}$ trials (**bounded** above and below)
  - Bernoulli for binary = special case of binomial when n=1
  - $Prob(y_i = c) = \frac{n!}{c!(n-c)!} p^c (1-p)^{n-c}$

p = probability of 1



Mean = np

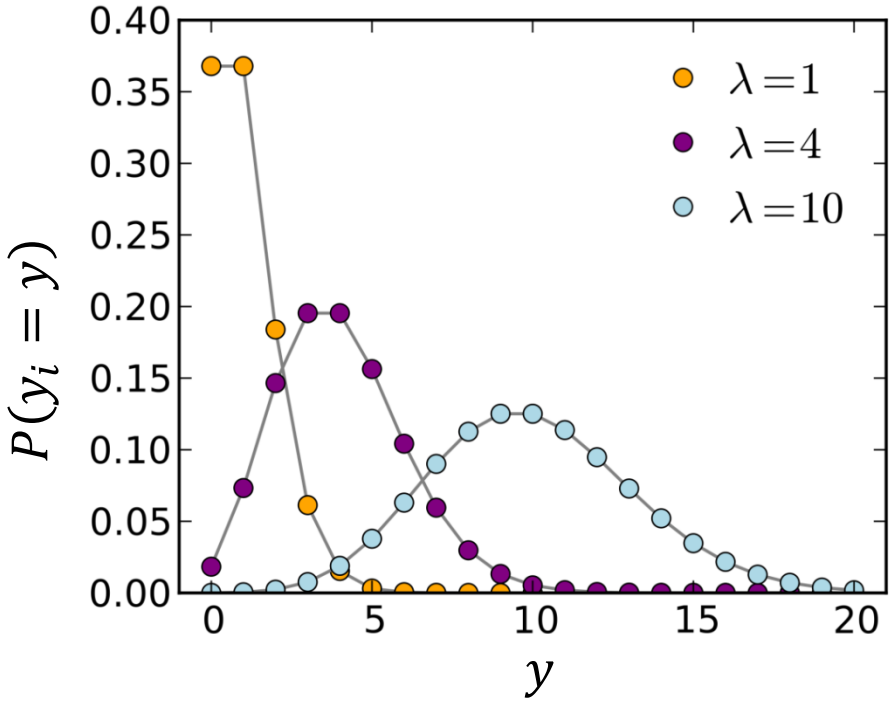
Variance = $np(1-p)$

Skewness = $\frac{1-2p}{\sqrt{np(1-p)}}$

Image borrowed from:
https://www.boost.org/doc/libs/1_42_0/libs/math/doc/sf_and_dist/html/math_toolkit/dist/dist_ref/dists/binomial_dist.html

Poisson Distribution: **Counts**

- **Poisson distribution has one parameter**, $\boldsymbol{\lambda}$, which is both its mean and its variance: λ = mean = variance; skewness = $\frac{1}{\sqrt{\lambda}}$
- $f(y_i|\lambda) = Prob(y_i = y) = \frac{\lambda^y * exp(-\lambda)}{y!}$ $y!$ is factorial of y

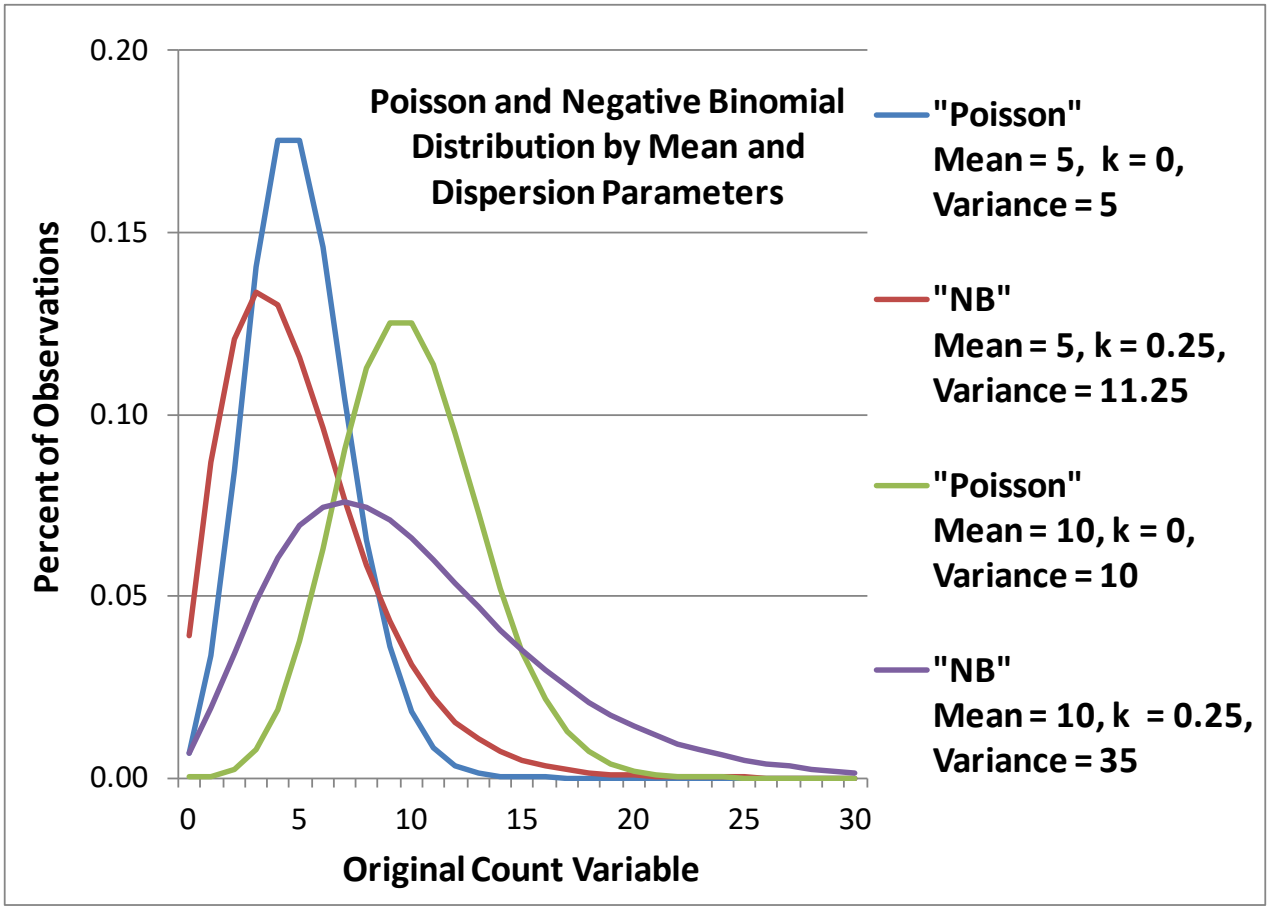


The dots indicate that only integer values are observed.

Distributions with a small λ are predicted to have a lot of 0's and will be asymmetric with positive skewness (tail off to the right). The larger the λ, the more symmetric the Poisson appears.

Image borrowed from: https://en.wikipedia.org/wiki/Poisson_distribution

Negative Binomial (NB) = “Stretchy” Poisson



$$\text{Mean} = \lambda$$
$$\text{Dispersion} = k$$

$$\text{Var}(y_i) = \lambda + k\lambda^2$$

A Negative Binomial model can be useful for **count outcomes with extra positive skewness**, but that otherwise follow a Poisson distribution.

Zero-Inflated Variables

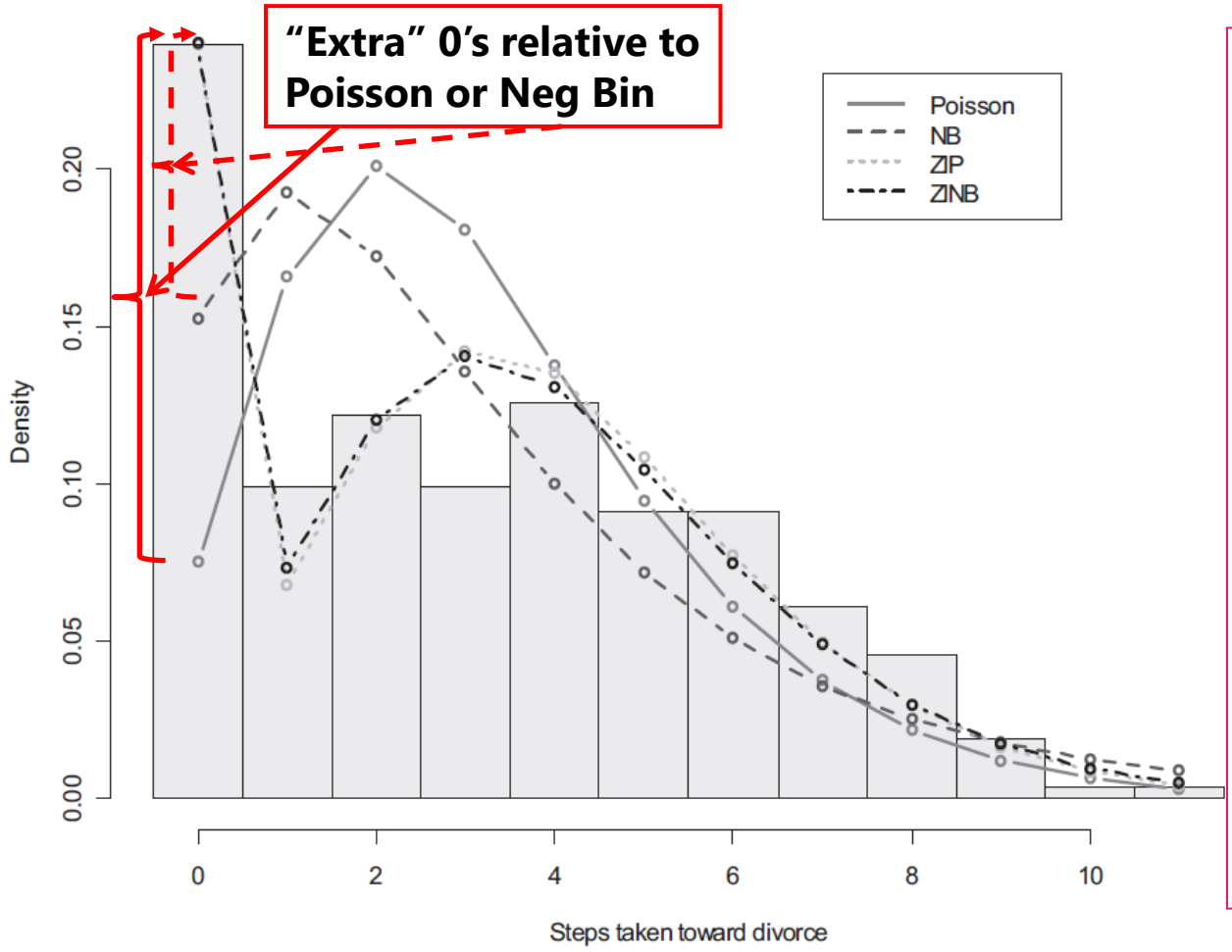


Figure 1. Histogram of Marital Status Inventory with predicted probabilities from regressions. NB = negative binomial; ZIP = zero-inflated Poisson; ZINB = zero-inflated negative binomial.

Zero-inflated distributions come in two flavors: Poisson (mean = variance) and Negative Binomial (variance exceeds mean).

When predictors have this type of distribution it can be helpful to think of them as **semi-continuous** in an "**if and how much**" model (my own terminology):

Pred1: =0 if x=0, 1 if x > 0
= Pred1 is binary
Pred2: =how much if x > 0
= Pred2 is quantitative

Image borrowed and doctored from: Atkins, D. C., & Gallop, R. J. (2007). Rethinking how family researchers model infrequent outcomes: A tutorial on count regression and zero-inflated models. *Journal of Family Psychology, 21*(4), 726-735.

B. Same Linear Predictive Model

- Your **outcome type** will likely guide you towards the most useful link function and conditional distribution
- Then you can include whatever **fixed effects** of predictors best address your study design and research questions, just s in GLMs estimated using ordinary least squares, with a few small differences:
  - The **specific names** that distinguish models with categorical from quantitative x_i predictors **are gone** from now on
  - They will be interpreted as **predicting the link-transformed** conditional mean (e.g., the logit of the probability; the log of the expected count)
  - ***F*-values** will show up without sums of squares and mean squares, but they are **interpreted the same way** (significance of multiple fixed slopes at once; weighted ratio of *known* to *unknown* info)
  - All parameters (fixed effects and variance-related terms) and their SEs will result from maximum likelihood estimation, but whether or not they are tested **using denominator DF will vary by software**
  - The use of conditional distributions without a separately estimated residual variance means that a **traditional R^2 will not be possible**

History of Generalized Linear Models

- Before ML estimation was widely available, other approaches were used to "handle" non-normality and non-constant variance; these should now all be considered as last resorts!
  - **Data transformations** (i.e., data cleaning, *shudder*)
    - e.g., positively skewed outcomes could be transformed via the square root or natural log to better approximate normality
    - e.g., an arc-sine transformation "stabilizes variance" (makes variance more constant) for proportions
    - e.g., a logit-transform creates an S-shaped curve to respect boundaries of predicted proportions in linear models
  - **Nonparametric statistics**: most are less flexible than generalized models because they still require some kind of least squares estimation
    - e.g., they may require rank transformations first (as in Spearman correlation) that throw away information about absolute distances
    - e.g., the same type of non-normal distribution must hold across groups

From Univariate to Multivariate

- This course will begin with prediction using generalized linear models of all kinds of outcomes, one at a time, but many types of data and/or research questions require multivariate models:
- **When y_i is still a single outcome conceptually, but:**
  - You have more than one outcome per person created by multiple conditions (e.g., longitudinal or repeated measures designs)
  - When your outcome is measured multiple times for a pair or group with distinguishable members (e.g., dyadic or family data)
- **When your hypotheses involve more than one y_i:**
  - To compare predictor effect sizes across outcomes (e.g., is the treatment effect bigger on outcome A than outcome B?)
  - You want to test indirect effects among them (i.e., mediation), so that a single variable is both a predictor and an outcome

From Univariate to Multivariate

- Ordinary least squares (OLS) has a "closed form" solution (its "sums of squares" formulae) when used for GLM for single outcomes
- For GLM for multiple outcomes, **OLS quickly becomes useless**…
  - Cannot handle missing outcomes (listwise-deletes entire person instead)
  - Only two options for modeling residual correlation between outcomes
  - Requires balanced data (same number of outcomes per higher unit)

- We will continue using maximum likelihood (ML) estimation for **multivariate models** to solve these problems, but some multivariate model variants will **require a switch in software**
  - Models in which *all variables are <u>either</u> predictors <u>or</u> outcomes* can be done by tricking univariate (regression-type) ML software (e.g., MIXED)
  - Otherwise, models must be estimated in ML using "truly" multivariate software (such as is used in path analysis or latent variable modeling)

Challenges in Truly Multivariate Software

- Trying to find ML one-size-fits-all multivariate software is tricky because you have to pay attention to the following options:
  - Whether **link functions** and alternative conditional distributions are available to predict non-normal outcomes
    - Whether it can use full-information estimation (uses all the data) or must use limited-information estimation (uses data summary)
  - Whether **predictors *must* or *can* be treated as outcomes** in order to allow persons with missing data to be included
    - If so, non-normal predictors must use non-normal distributions
  - Whether the **model can expand** to address other sources of person dependency (i.e., persons nested in multiple groups)

- To overcome these limitations, we *may* have to transfer to M*plus* software (i.e., instead of SAS, Stata, SPSS, or R)

This Course and Beyond: Lego #4

- **This course** will help you understand how to **combine**:
  - (1) **linear models** and (3) **link functions** to predict any kind of outcome, which is possible through (2) the use of **ML estimation** (as well as how it is used to assess fit in multivariate models)

- Conquering this material serves two distinct purposes:
  - Being able to **predict any kind of outcome** in order to test univariate or multivariate hypotheses is useful in and of itself!
  - In addition, these are all **essential pre-requisite** skills that I usually must review or teach from scratch in advanced classes that also include **Lego #4: random effects and/or latent variables**
    - Multilevel models, structural equation models (SEM), multilevel SEM

- So please stick with me—either way, I hope you won't regret it! ☺