Biostatistics 140.754
Advanced Methods in Biostatistics IV

Jeffrey Leek

Assistant Professor
Department of Biostatistics
jleek@jhsph.edu

Course Information

- Welcome.
- The primary focus of this course is regression modeling, along with other more "modern" approaches for estimating or predicting the relationship between random variables.
- The prerequisites for this course are Biostatistics 140.751-140.753.
- All learning outcomes, syllabus, motivation, grading, etc. are available from the course website: www.biostat.jhsph.edu/~jleek/teaching/2011/574
- Lecture notes will be posted the night before class.
- Course evaluation will consist of a weekly reading assignment, a biweekly homework assignment, and a final project.

Course Information - Credits

Ken Rice (UW) - (slides with a † are directly lifted from hi

Jon Wakefield (UW)

Brian Caffo (JHU)

+ Assorted others as mentioned in the text. Any mistakes, typos, or otherwise misleading information is "measurement error" due to me.

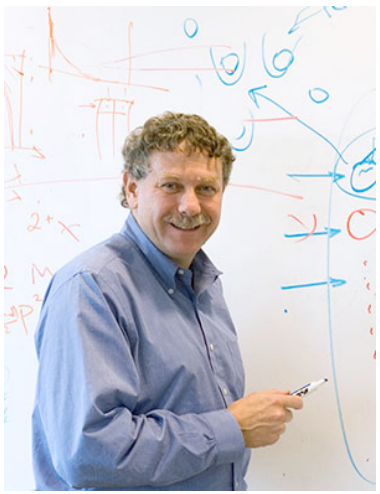
What's So Great About Applied Statistics?

"I keep saying the sexy job in the next ten years will be statisticians. People think I'm joking, but who would've guessed that computer engineers would've been the sexy job of the 1990s?"

- Hal Valarian (Google Chief Economist)

“Applied Statisticians”

Eric Lander



Director – Broad

Steven Levitt

“Freak-onomics”

Nate Silver

fivethirtyeight.com

Daryl Morey

Houston Rockets GM

"Jobs For Applied Statisticians"

Groupon Jobs

Current Openings > Data Scientist

Data Scientist

Engineering | Palo Alto, CA, United States

Apply Now

Send Jobvite

Preferred Skills:

1. PhD in data mining, machine learning, statistical analysis, applied mathematics or equivalent.
2. 3+ years hands-on practical experience with large scale data analysis
3. Fluency in analytical tools such as SAS, R, etc.

Google

Home › About Google › Jobs › US locations › California - Mountain View › Marketing & Communications › Marketing ›

Jobs
Joining Google
Life at Google
Office locations

Quantitative Analyst - Mountain View

This position is based in New York, NY., Mountain View, CA. or Boulder, CO.

Requirements:

- PhD in Statistics or Econometrics required or equivalent work experience.
- Experience with R/Splus.
- Coursework in Bayesian methods, longitudinal analysis, and experimental design desirable.
- Experience with Python, Perl, SQL is desired, but not required.

Course Information - How does 574 fit in? †

574 is an advanced, Ph.D. level course. The following are assumed:

- **Linear algebra**; expressions like $(\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{Y}$ should make sense to you.
- **Introductory probability**; manipulation of distributions, Central Limit Theorem, Laws of Large Numbers, some likelihood theory
- **Introductory Regression**; some familiarity with multiple regression will be helpful
- **The R Language**; sufficient to implement the material above (and look up new stuff in help files)

Please note: much of 574 will interpret regression from a non-parametric point of view. This is a modern approach, and may differ from classical material you have seen elsewhere.

Course Information - How does 574 fit in?

574 is a **methods** course

- The main aim is to understand how/why methods work and what practical situations where they will be most useful.
- Formal math will be limited in the lecture notes (unlike in 673-674, 771-772), so expect some hand-waving (e.g. "...under mild regularity conditions").
- Many of the course example will be short/stylized. However, the goal of the course is to provide both understanding of specific methods and their implementation/application/interpretation.

Course Information - How does 574 fit in?†

The term "methods" is somewhat open to interpretation - this is one potential way to break journals down to give some insight

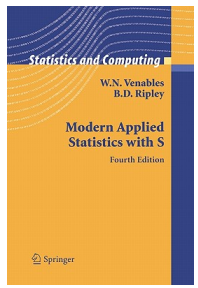
- **Theory:** Annals of Statistics, JRSSB, Statistica Sinica
- **Data Analysis:** JASA A&CS, JRSSC, Nature, NEJM, JAMA, Neuroimage, Genome Biology
- **Methods:** Biometrics, Annals of Applied Statistics, Biostatistics, Statistics in Medicine, Neuroimage, Genome Biology

Modern methods papers use simulation studies to illustrate statistical properties; we will often do the same.

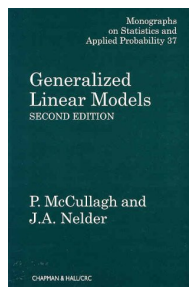
Most PhD theses "resemble" methods papers, and contain material similar to that discussed in 574. A focus of this course will be reading, understanding, and learning to construct academic papers.

Course Info - Textbooks

There is no fixed textbook for this course. A couple of useful books may be:



Modern Applied Statistics with S

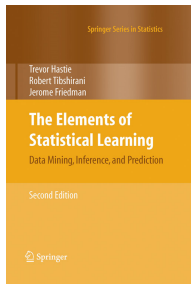


Generalized Linear Models

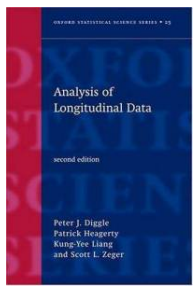
Research papers will be featured, for more recent topics - 574 is more cutting edge than some other courses we teach.

Course Info - Textbooks

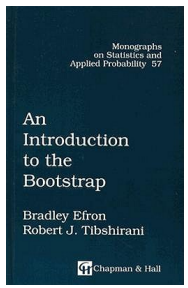
Another couple of "classics" applied statisticians should have access to:



Elements of Statistical Learning



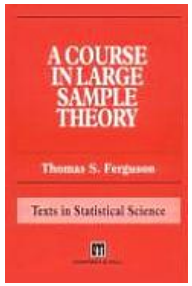
Analysis of Longitudinal Data



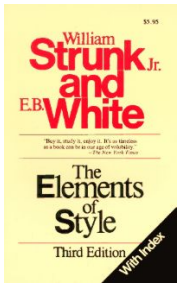
An Introduction to the Bootstrap

More Ridiculously Useful Books

Another couple of really useful books - not 100% related to course content, but **highly** recommended



A course in large sample theory [1]



The Elements of Style

http://www.biostat.jhsph.edu/~jleek/teaching/2011/754/reading/StrunkandWhite.pdf

[1] The instructor's favorite statistics book

Course Info - Course Content

- Review of ideas behind regression
- Non-parametric inference (generalized method of moments)
- Likelihood + Quasi-Likelihood inference
- Bayesian inference
- Analysis of correlated data - generalized estimating equations
- Bootstrapping
- Model selection/shrinkage (Lasso, etc.)
- Factor analysis/principal components analysis
- Interaction-based approaches to prediction/association (i.e. CART)
- Multiple testing

Outline of Today's Lecture

- Background (randomness, parameters, regression)
- Regression with estimating equations
- Sandwich estimators of variance

Terminology†

- The response variable will be termed the outcome. Usually we wish to relate the outcome to covariates.

Abbreviation	Y	X (or Z, U)
Preferred name[2]	Outcome	Covariate(s)
Other names:	Response Output Endpoint	Regressors, Predictors Input Explanatory Variable
Confusing Name	Dependent	Independent

- Predictor has causal connotations. [In]dependent is a poor choice (the covariates need not be independent of each other - and may be *fixed*, by an experimenter)
- In 574 we consider Y and X which are continuous, categorical, or counts; later in the course multivariate outcomes are briefly considered (more on that in 755/56). Outcomes which are censored or mixed (e.g. alcohol consumption) are also possible. Categorical variables may be nominal or ordinal.

[2] *Preferred by me*

What is Randomness?†

You may be used to thinking of the stochastic parts of random variables as just chance. In **very select** situations this is fine; radioactive decay really does appear to be just chance [3]

However, this is not what random variables actually represent in most applications, and it can be a **misleading** simplication to think that its just chance that prevents us knowing the truth.

To see this, consider the following thought experiments...

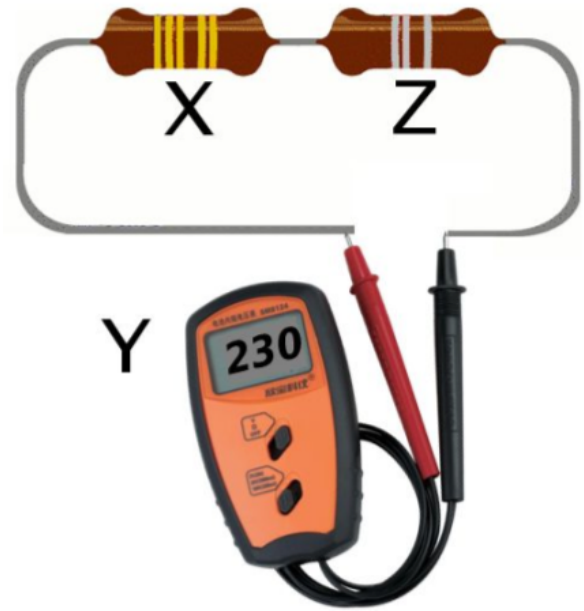
[3] *But ask Brian Caffo about this...*

What is Randomness?†

Recall high school physics... For two resistors "in series", the resistances are added to give a total (Y , measured in Ohms, Ω) which we record **without error**

We know the number of gold stripes (X) and silver stripes (Z). We also know that each resistance is $\propto$ number of stripes.

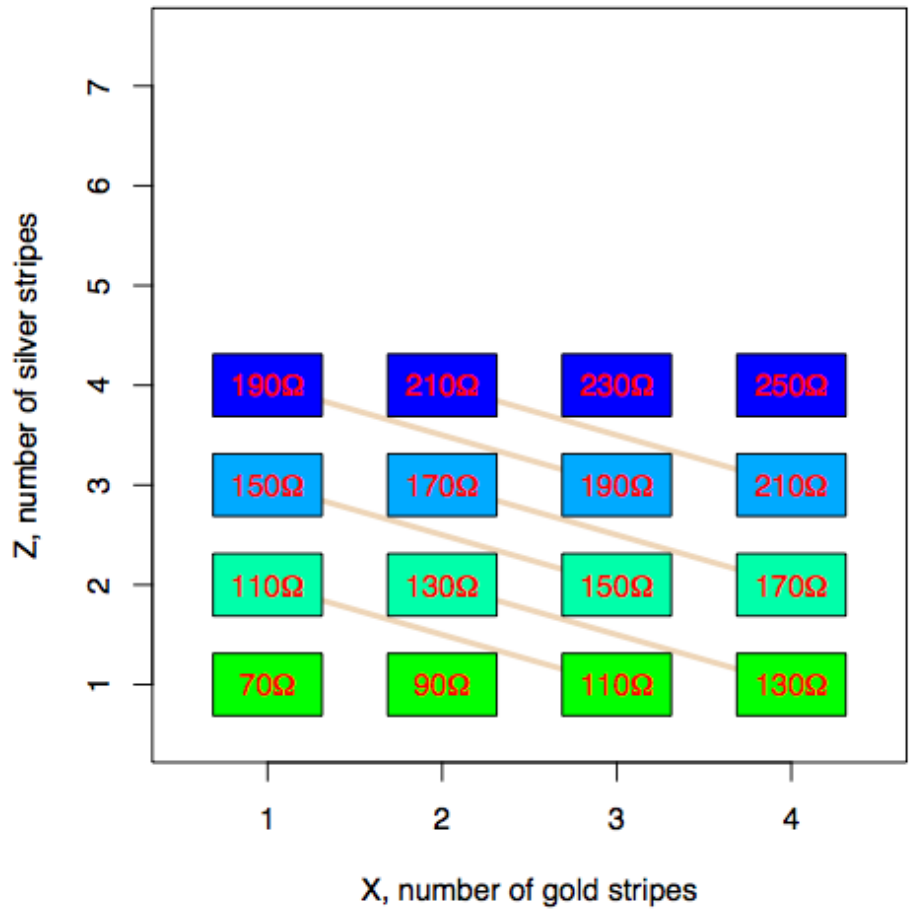
Q. How much resistance do stripes of each color correspond to?



What is Randomness?†

Thought experiment #1; Note that in this situation there **no** "measurement error" or "noise", and **nothing random** is going on.

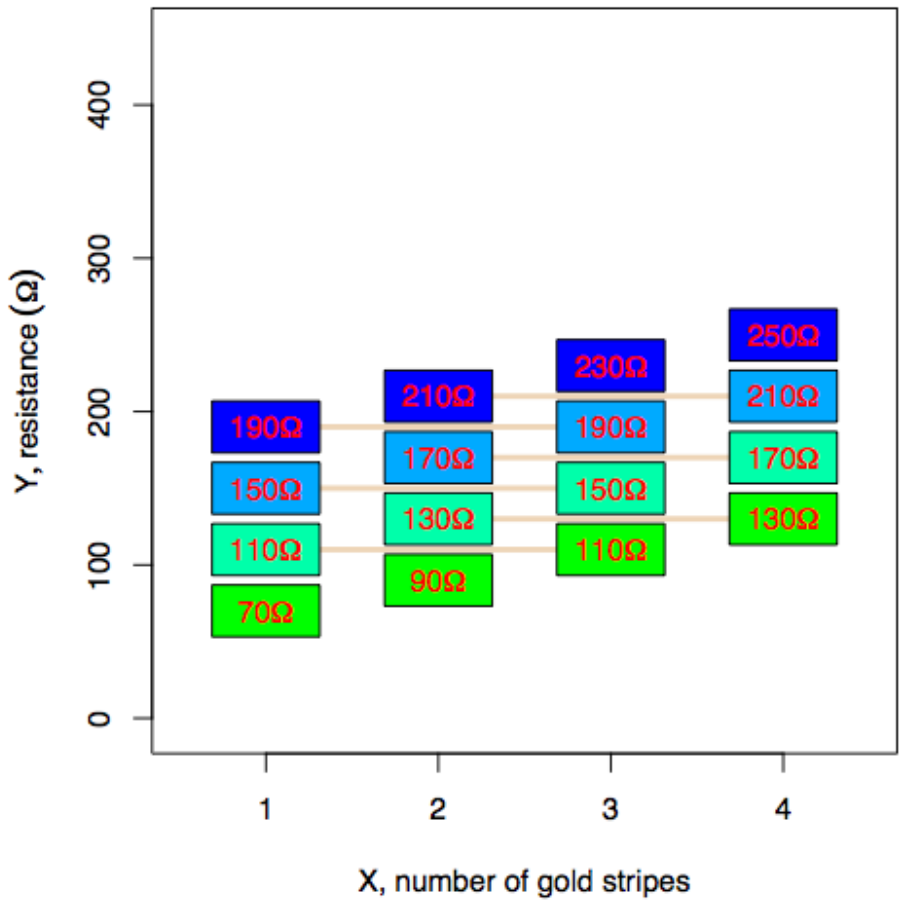
What is the "value" of each goldstripe?



What is Randomness?†

Thought experiment #1; Note that in this situation there **no** "measurement error" or "noise", and **nothing random** is going on.

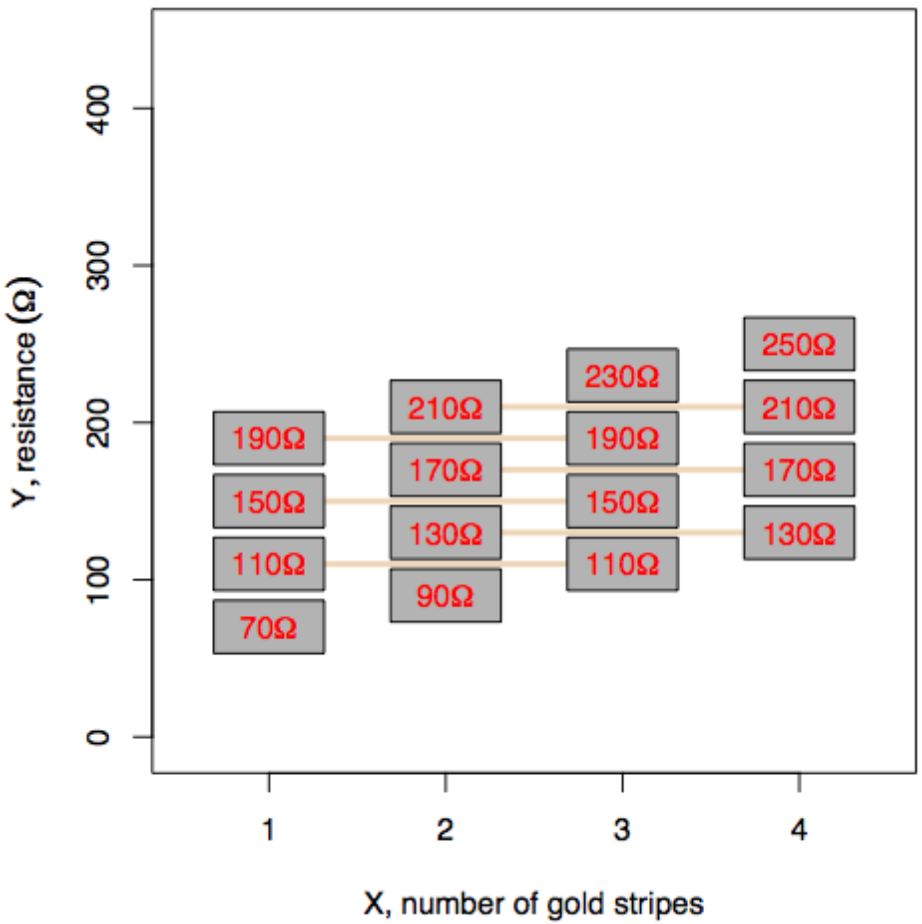
What is the difference between X and X+1?



What is Randomness?†

Thought experiment #1; Note that in this situation there **no** "measurement error" or "noise", and **nothing random** is going on.

What is the difference between X and X+1?



Thought Experiment Math†

Here's the truth;

$$\mathbf{Y}_{n\times 1} = \gamma_0 \mathbf{1}_{n\times 1} + \gamma_1 \mathbf{X}_{n\times 1} + \gamma_2 \mathbf{Z}_{n\times 1}$$

where n is evenly distributed between all X, Z combinations.

But not knowing Z, we will fit the relationship

$$\mathbf{Y} \approx \beta_0 \mathbf{1} + \beta_1 \mathbf{X}$$

Here "fit" means that we will find **e orthogonal** to $\mathbf{1}$ and $\mathbf{X}$ such that

$$\mathbf{Y} = \beta_0 \mathbf{1} + \beta_1 \mathbf{X} + \mathbf{e}$$

By linear algebra (i.e. projection onto $\mathbf{1}$ and $\mathbf{X}$) we must have

$$\mathbf{e} = \mathbf{Y} - \left(\frac{\mathbf{Y}\cdot\mathbf{1}}{n} - \frac{\mathbf{Y}\cdot(\mathbf{X}-\bar{\mathbf{X}}\mathbf{1})}{(\mathbf{X}-\bar{\mathbf{X}}\mathbf{1})\cdot(\mathbf{X}-\bar{\mathbf{X}}\mathbf{1})}\mathbf{X} \right)\mathbf{1} - \left(\frac{\mathbf{Y}\cdot(\mathbf{X}-\bar{\mathbf{X}}\mathbf{1})}{(\mathbf{X}-\bar{\mathbf{X}}\mathbf{1})\cdot(\mathbf{X}-\bar{\mathbf{X}}\mathbf{1})} \right)\mathbf{X}$$

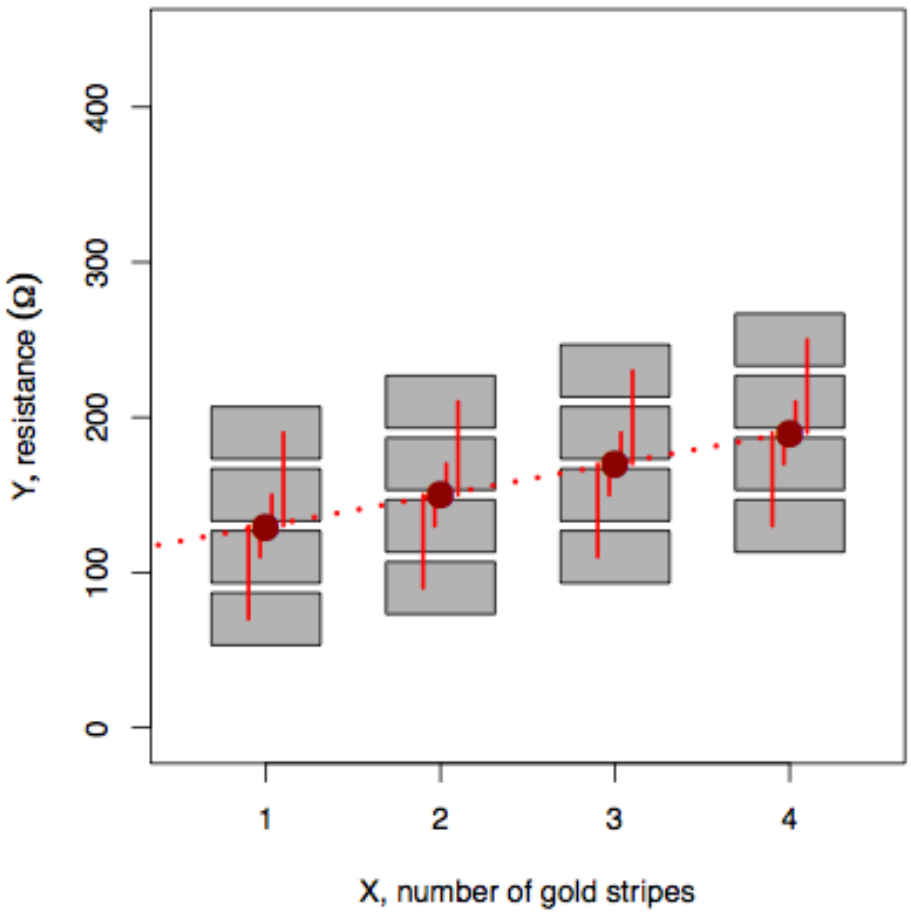
where $\bar{\mathbf{X}} = \mathbf{X}\cdot\mathbf{1}/(\mathbf{1}\cdot\mathbf{1}) = \mathbf{X}\cdot\mathbf{1}/n$, i.e. the mean of $\mathbf{X}$ - a scalar.

Thought Experiment Math?†

The fitted line, with $\mathbf{e}$

Note the orthogonality to $\mathbf{1}$ and $\mathbf{X}$

What's the slope of the line?



Thought Experiment Math?†

What to remember (in "real" experiments too);

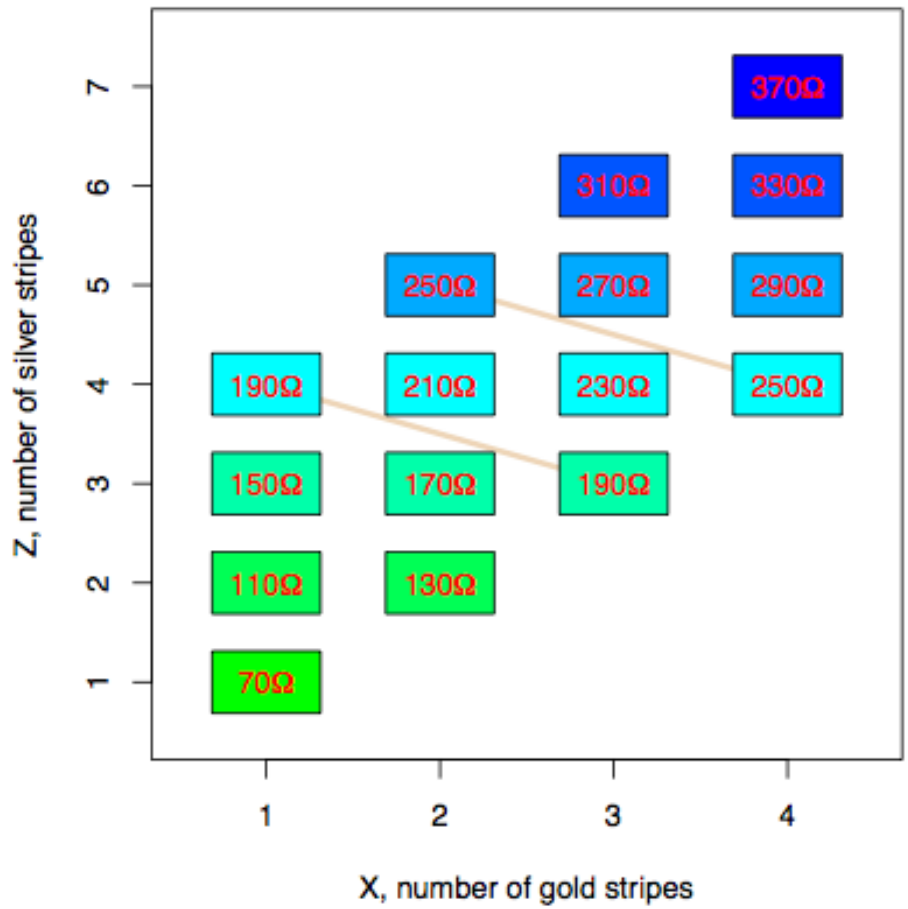
- The "errors" **represent** everything that we didn't measure.
- **Nothing** is random here - we just have imperfect information
- If you are *never* going to know Z (or can't assume you know a lot about it) this sort of "marginal" relationship is all that *can* be learned

What you *didn't* measure can't be ignored...

Thought Experiment #2 †

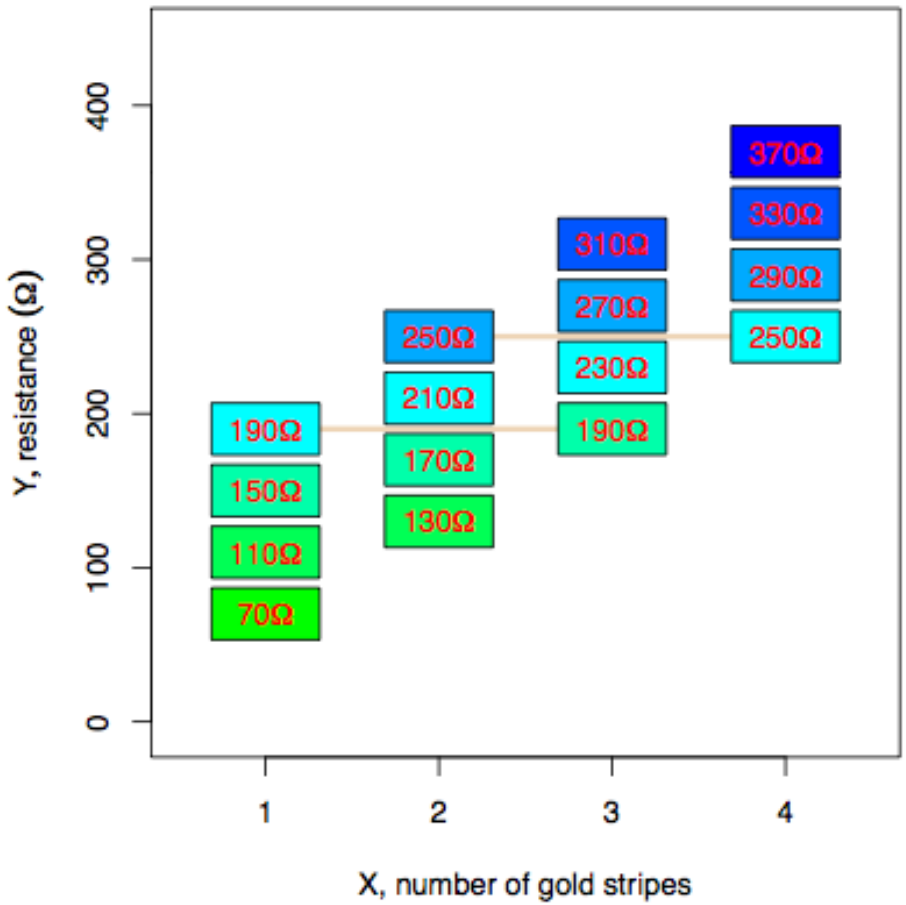
A different "design"

What is going on?



Thought Experiment #2 †

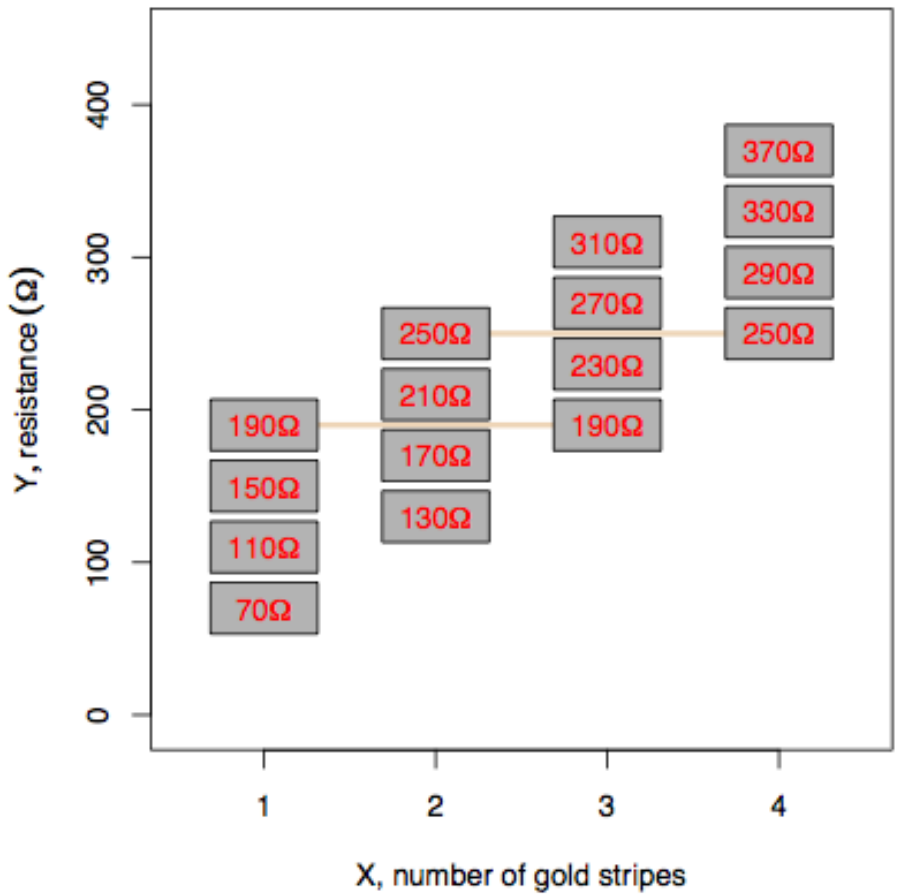
Plotting Y against X;



Thought Experiment #2 †

Plotting Y against X;

... and not knowing Z

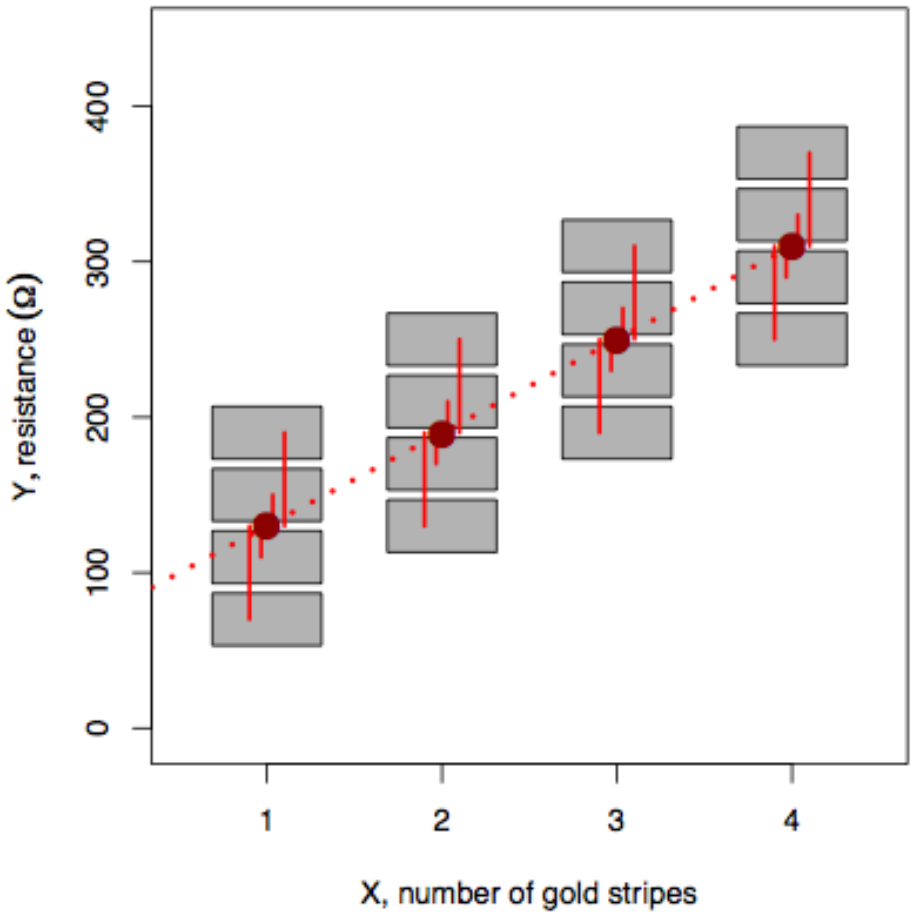


Thought Experiment #2 †

Here's the fitted line;

... what's the slope?

What would you conclude?



Thought Experiment #2 †

Here's the truth, for both $\mathbf{Y}$ and $\mathbf{Z}$;

$$\begin{aligned} \mathbf{Y} &= \gamma_0 \mathbf{1} + \gamma_1 \mathbf{X} + \gamma_2 \mathbf{Z} \\ \mathbf{Z} &= \theta_0 \mathbf{1} + \theta_1 \mathbf{X} + \boldsymbol{\epsilon} \end{aligned}$$

where $\boldsymbol{\epsilon}$ is orthongal to $\mathbf{1}$, $\mathbf{X}$. Therefore,

$$\begin{aligned} \mathbf{Y} &= \gamma_0 + \gamma_1 \mathbf{X} + \gamma_2(\theta_0 + \theta_1 \mathbf{X} + \boldsymbol{\epsilon}) \\ &= (\gamma_0 + \gamma_2\theta_0)\mathbf{1} + (\gamma_1 + \gamma_2\theta_1)\mathbf{X} + \gamma_2\boldsymbol{\epsilon} \\ &\equiv \beta_0 \mathbf{1} + \beta_1 \mathbf{X} + \mathbf{e} \end{aligned}$$

and we get $\beta_1 = \gamma_1$ if (and only if) there's "nothing going on" between Z and X. The change we saw in the $Y - X$ slope (from #1 to #2) follows exactly this pattern.

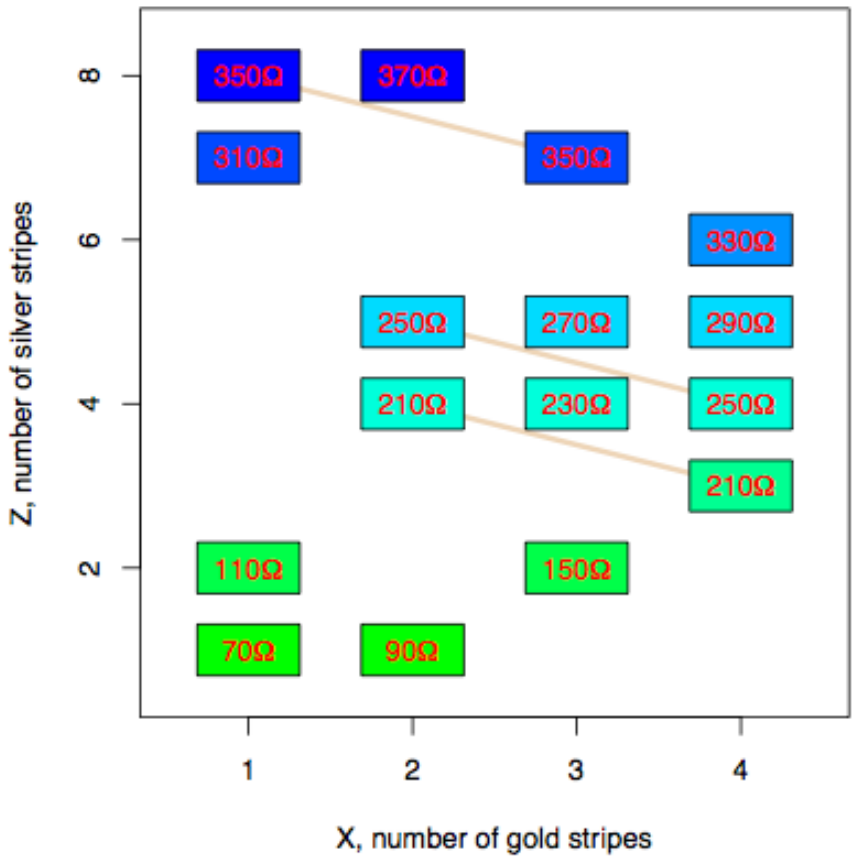
Thought Experiment #2 †

- The marginal slope β_1 is not the "wrong" answer, but it may not be the same as γ_1.
- Which do you wnat? The $Y - Z$ slope if Z is fixed or if Z varies with X in the same way it did in your experiment?
- No one needs to know that Y is being measured for $\beta_1 \neq \gamma_1$ to occur.
- The "observed" $\mathbf{e}$ are actually $\gamma_2 \boldsymbol{\epsilon}$ here, so the "noise" doesn't simply reflect the $Z - X$ relationship *alone*

Thought Experiment #3 †

A final "design"

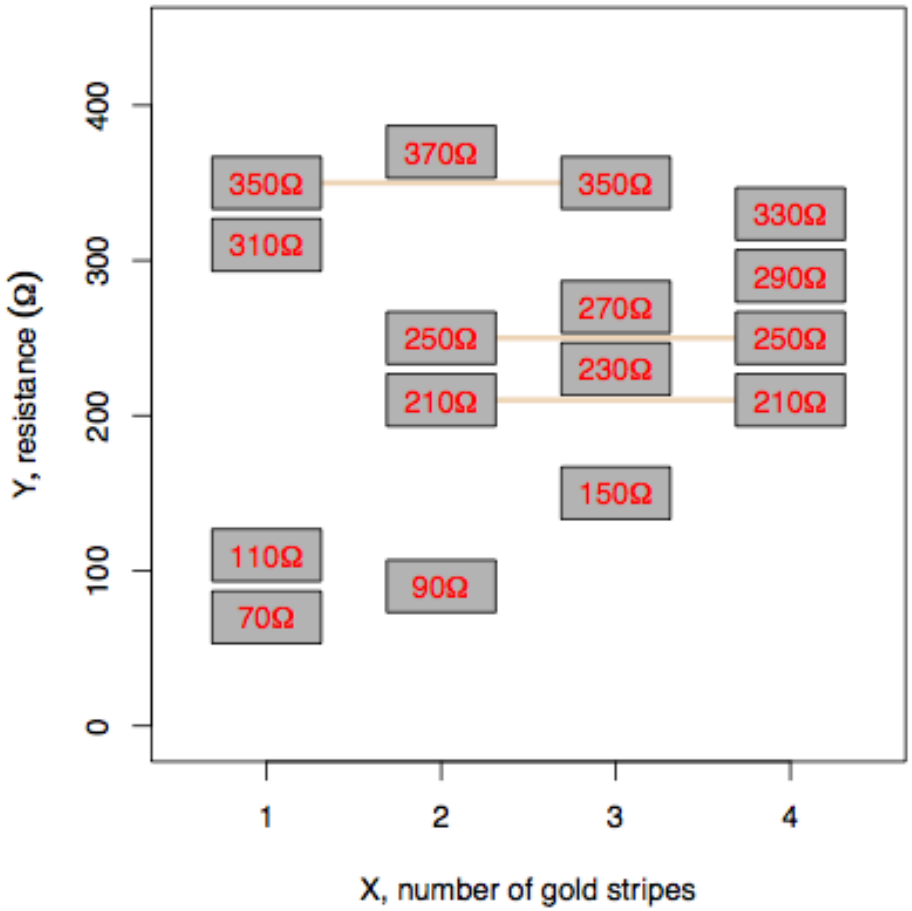
... a real mess!



Thought Experiment #3 †

A final "design"

... plotting Y vs. X

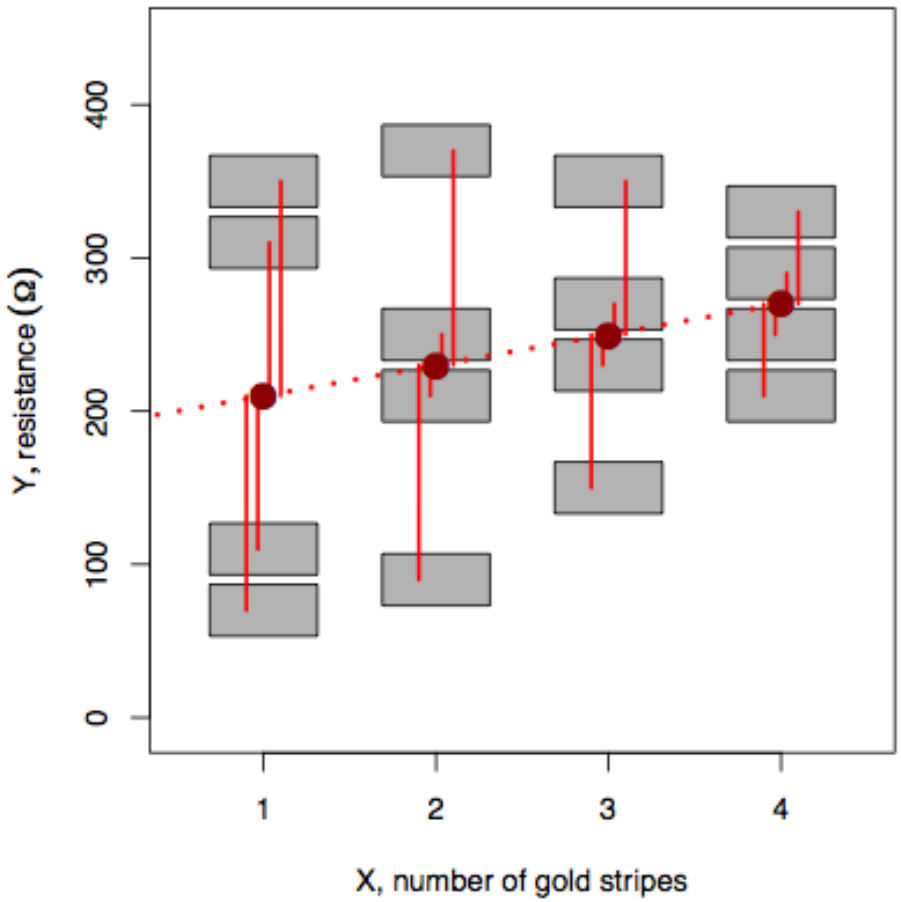


Thought Experiment #3 †

A final "design"

... plotting Y vs. X

(Starts to look like real data!)



Thought Experiment #3 †

- Z and X were orthogonal - what happened to the slope?
- *But* the variability of Z depended on X. What happened to **e**, compared to #1 and # 2? We can extend all these

  arguments to $\mathbf{X}_{n\times p}$ and $\mathbf{Z}_{n\times q}$ - see Jon Wakefield's book for more. Reality also tends to have > 1 "un-pretty" phenomena per situation!

  In general, the nature of what we call "randomness" depends **heavily** on what is going on unobserved. Its only in extremely simple situations[4] that unobserved *patterns* can be dismissed without careful thought. In **some** complex situations they *can* be dismissed, but only after careful thought.

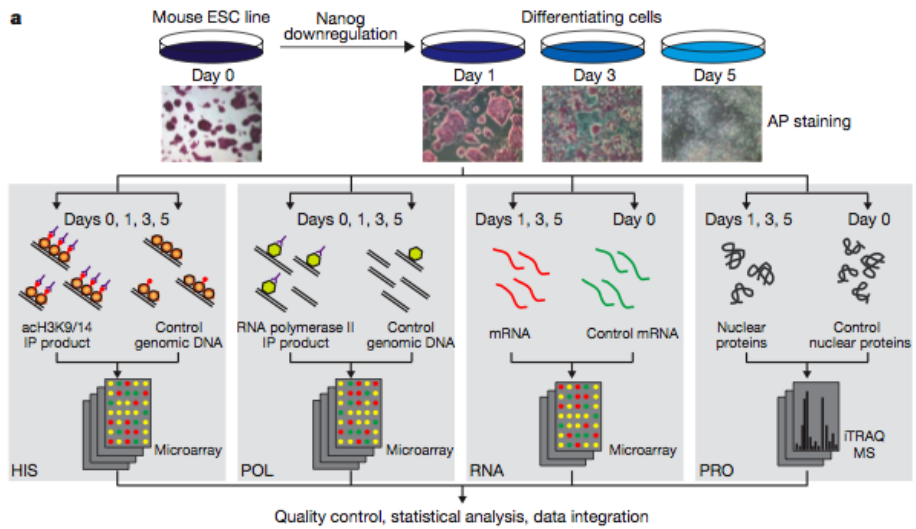
[4] ...which probably don't require a PhD statistician

Reality Check †

This is a realistically- complex "system" you might see in practice

Your "X" might be time (developmental) and "Y" expression of a particular gene

Knowing the Y-X relationship is clearly useful, but pretending that all the Z -X relationships are pretty is naïve (at best)



Reality Check †

With reasonable sample size n, inference (i.e. learning about β) is possible without making strong assumptions about the distribution of Y , and how it varies with X. It seems prudent to avoid these assumptions as "modern" approaches do.

- If you have good a priori reasons to believe them, distributional assumptions may be okay and may help substantially
- For small n this may be the only viable approach (other than quitting)
- For tasks other than inference (e.g. prediction) assumptions may be needed.
- Checking distributional assumptions after you've used them doesnt actually work very well. Asking the data "was I right to trust you just now" ? or "did you behave in the way I hope you did?" is not reliable, in general.

Reality Check †

If you have to start making distributional assumptions:

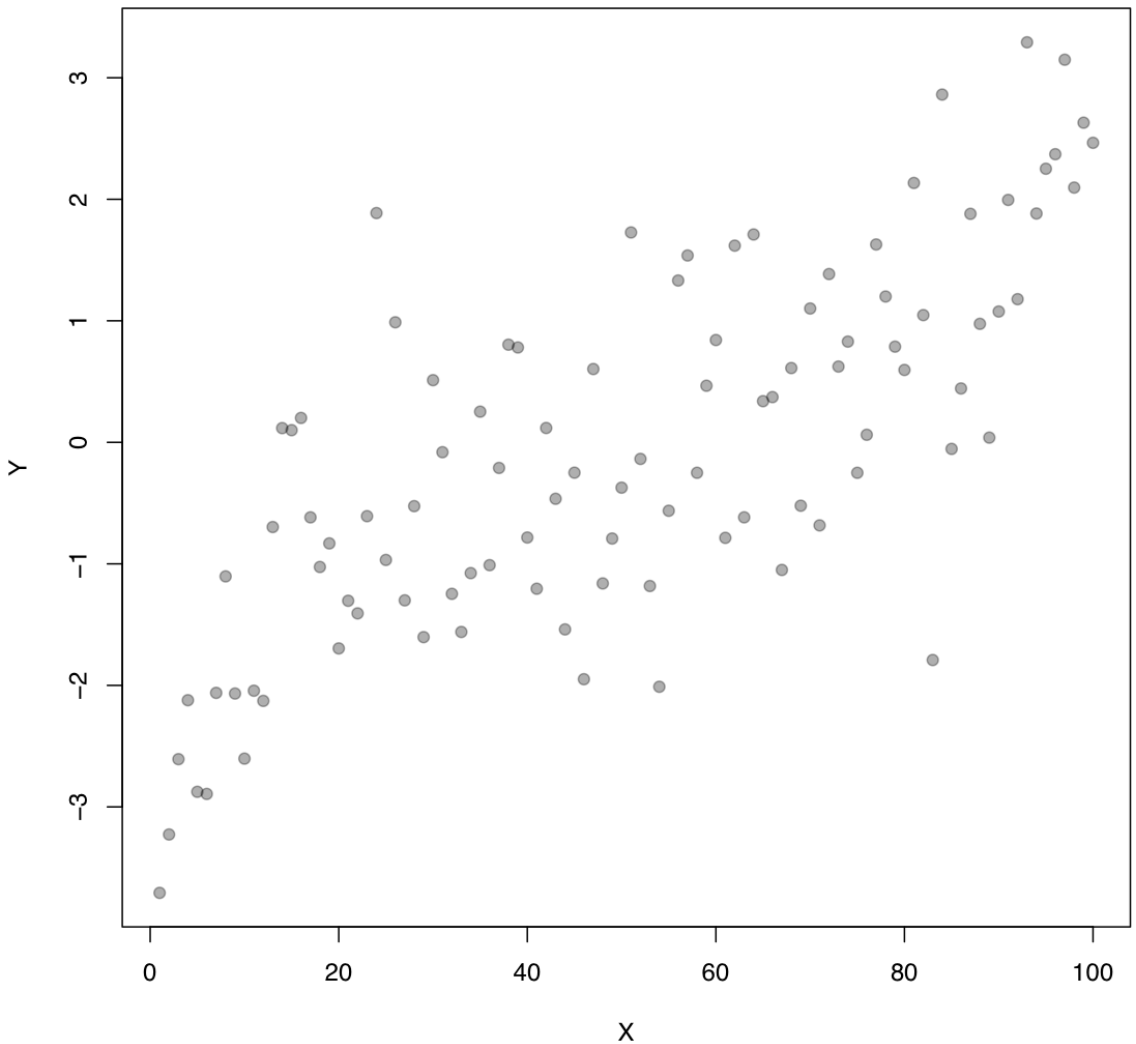
- Adding lots of little effects $\longrightarrow$ Normal distributions
- Binary events $\longrightarrow$ Bernoulli, and Binomial
- Counting lots of rare events $\longrightarrow$ Poisson
- Continual (small) hazard of an event $\longrightarrow$ Weibull

... but note these are rather stylized, minor modications break them, e.g. different event rates $\longrightarrow$ overdispersed Poisson.

However, methods which use classical assumptions often have other interpretations. For example, using $\bar{Y}$ (the sample mean) as an estimator can be motivated with Normality, but we don't need this assumption in order to use Y.

What is a parameter?†

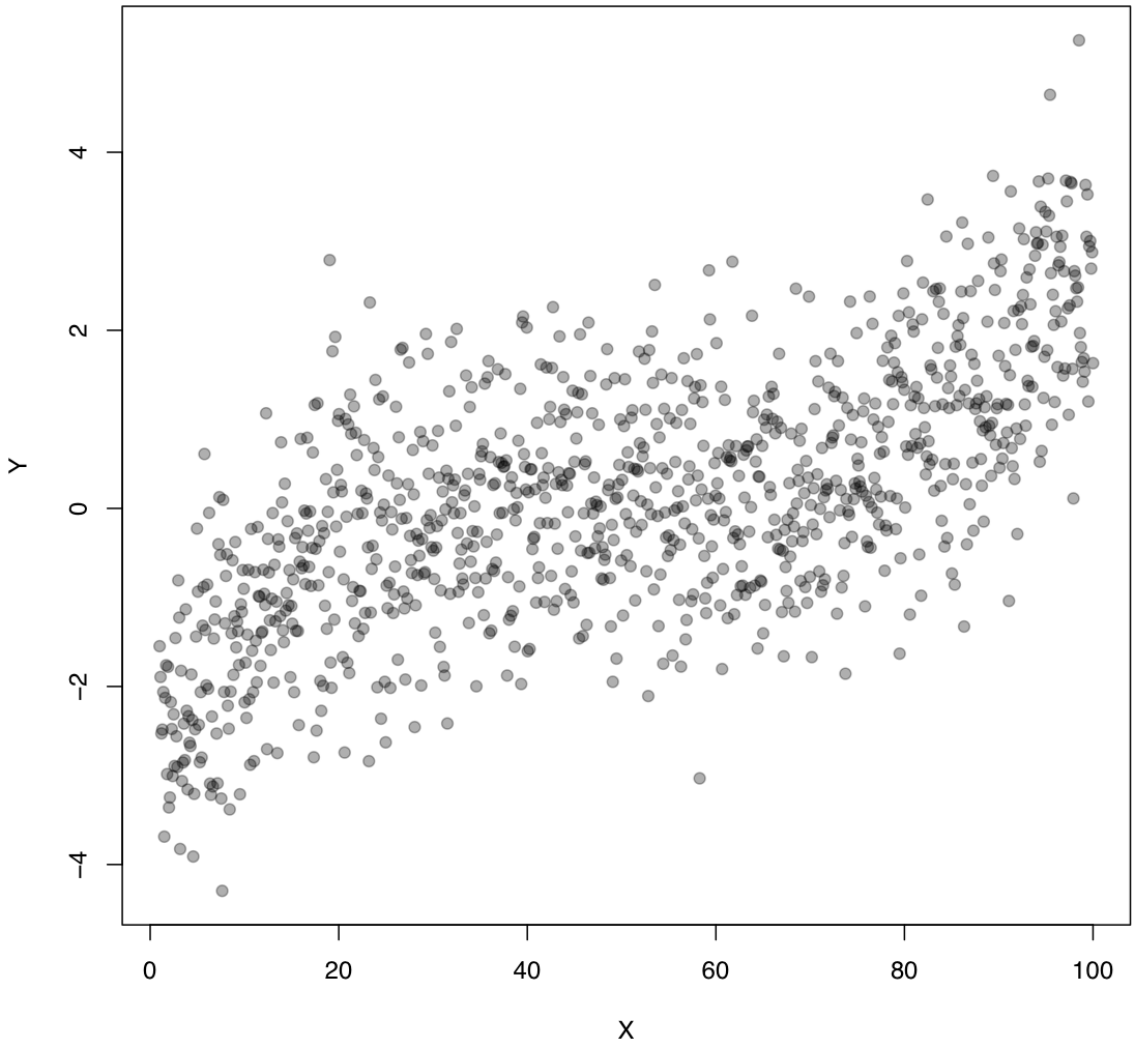
From previous courses you will be used to this kind of plot



... and also used to "manipulating" the sample in several ways

What is a parameter?†

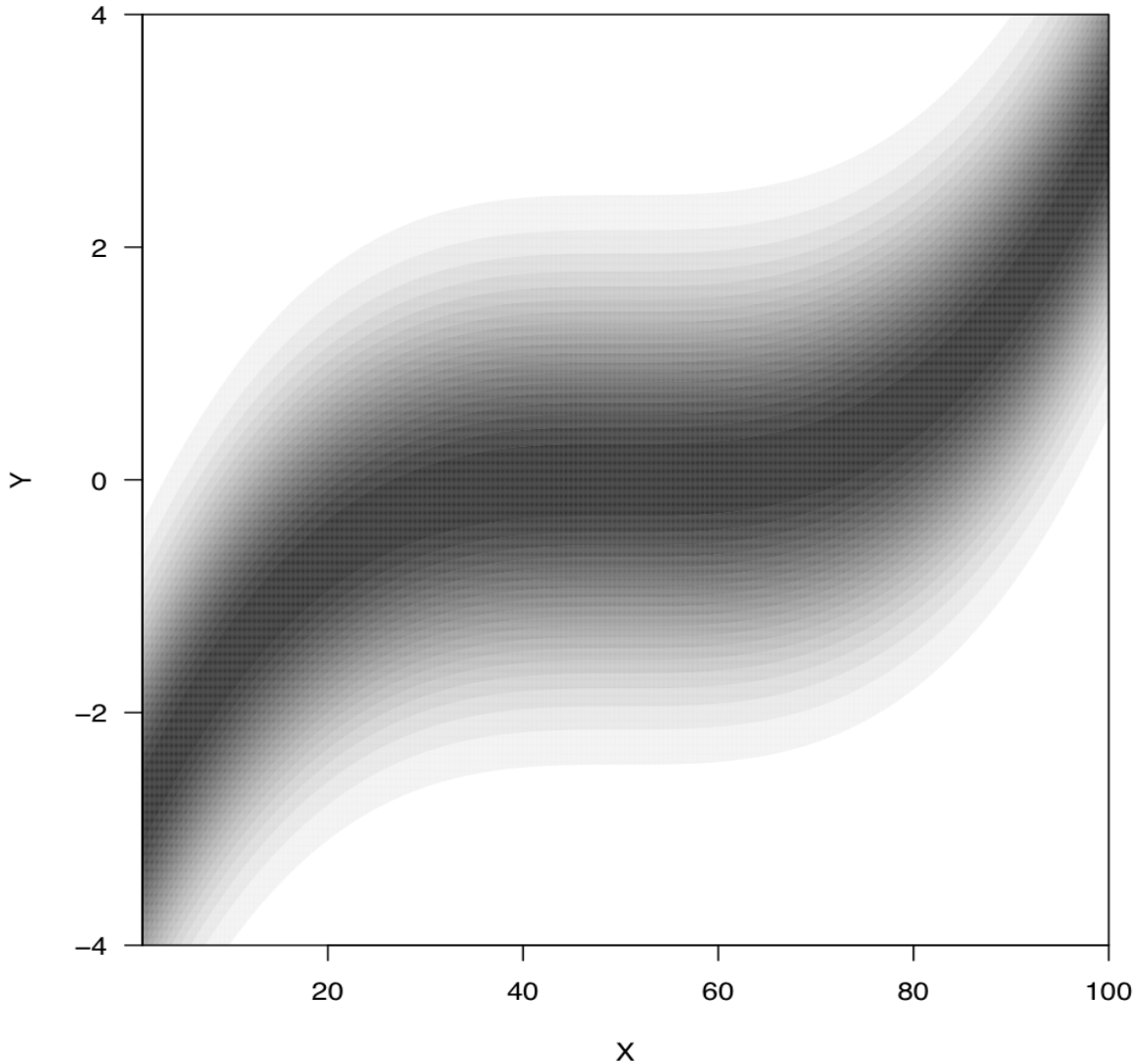
You may have seen larger sample sizes,



... this sample can also be "manipulated"

What is a parameter?†

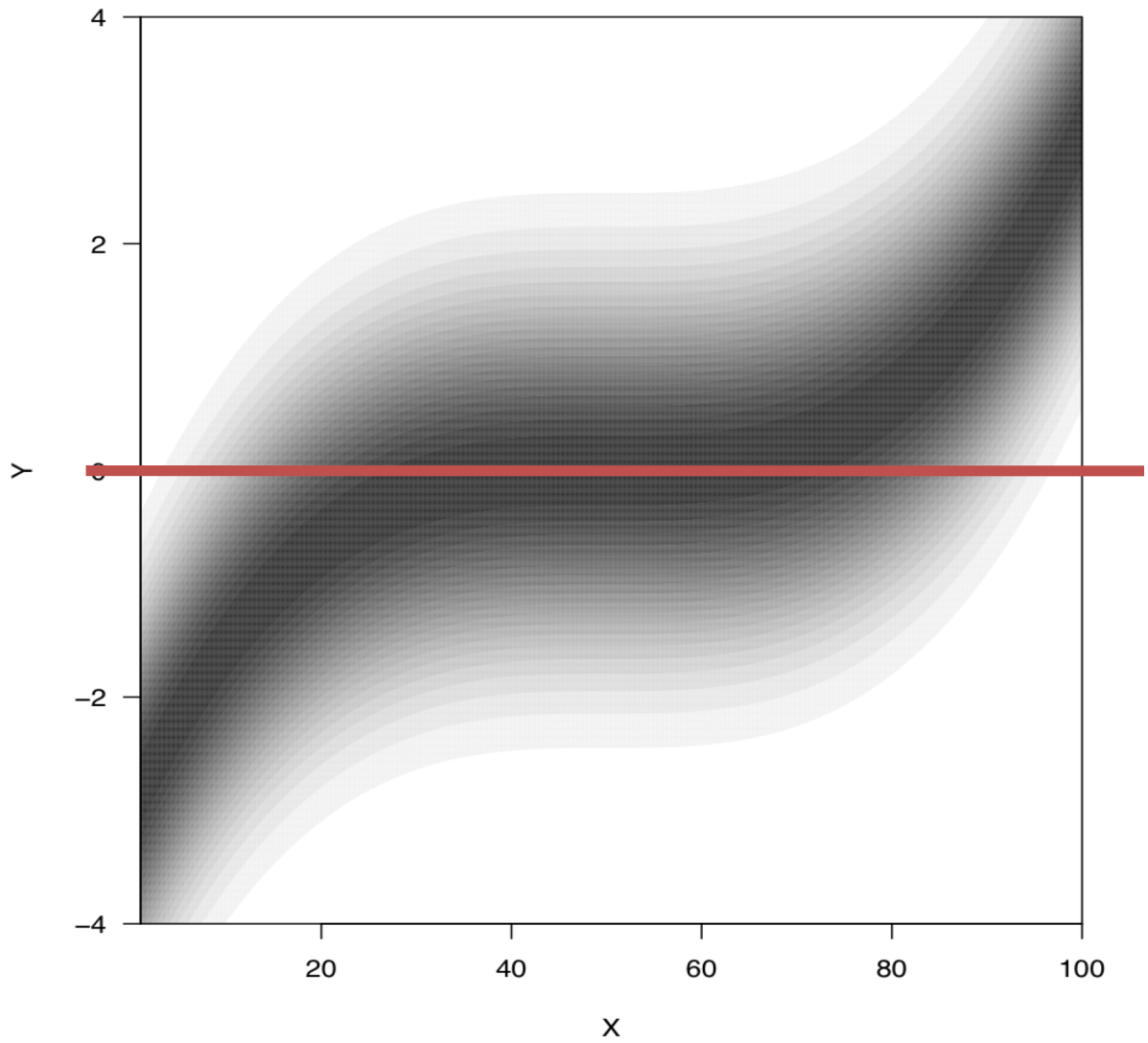
To define parameters, think of an **infinite** "super"-population;



... and consider (simple) ways to manipulate what we see;

What is a parameter?†

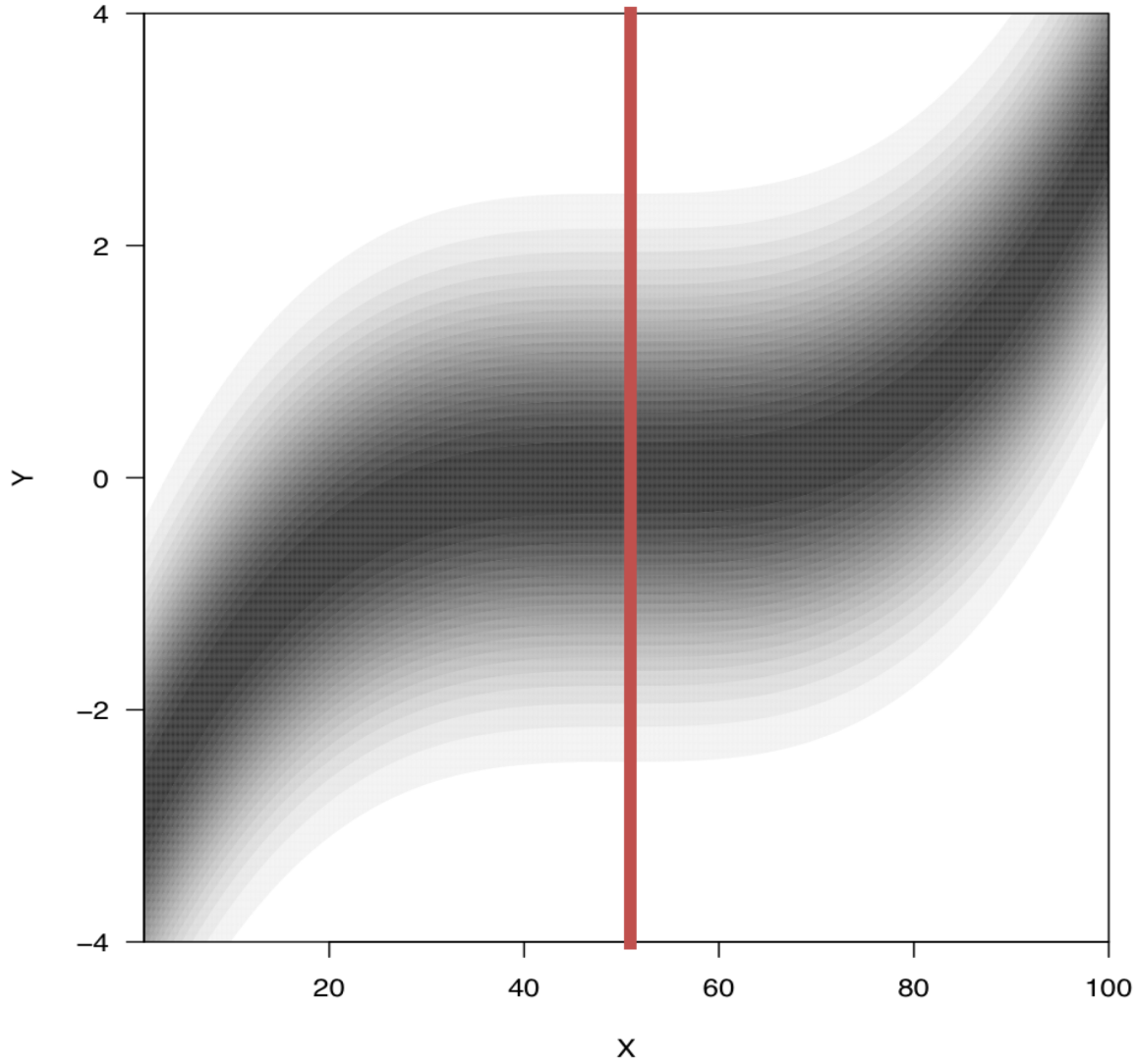
The mean of X;



(note: requires finite moments of X to be well-defined)

What is a parameter?†

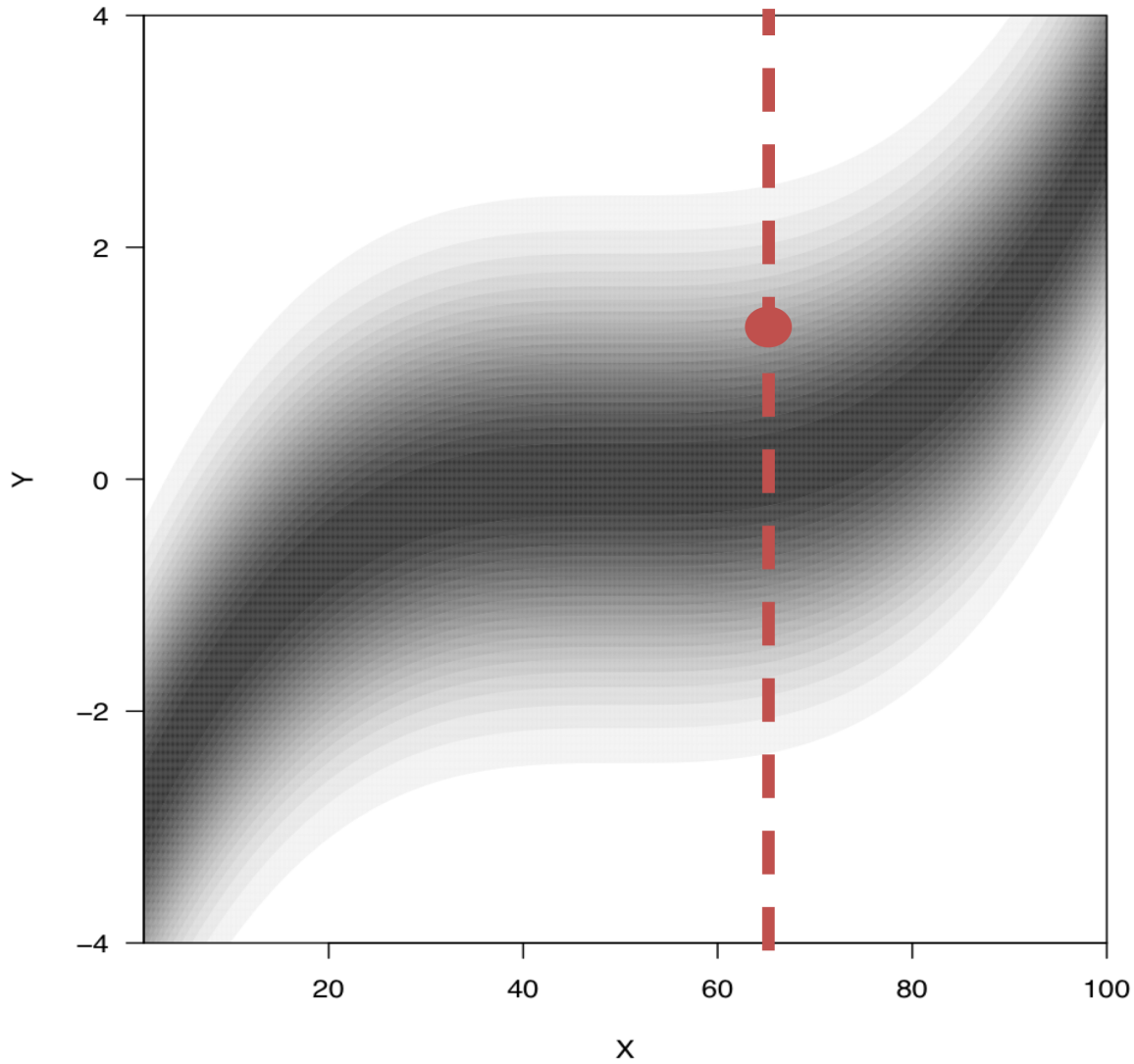
The mean of Y ;



... *mild* regularity conditions also apply

What is a parameter?†

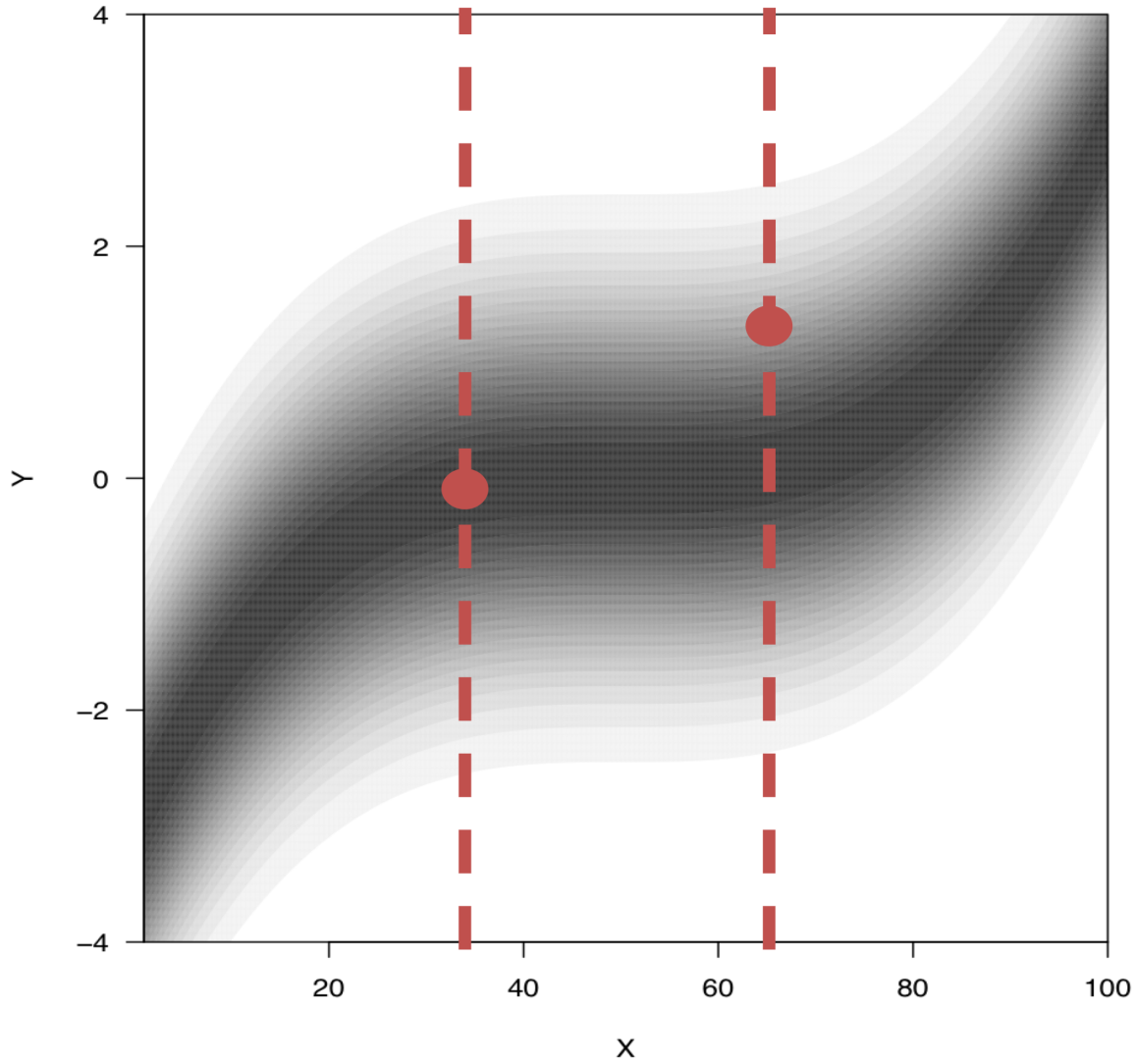
The mean of Y at a given value of X



... only sensible if you know the given value of X (!)

What is a parameter?†

Difference in mean of Y , between two values of X;



... which is unchanged, if $Y \rightarrow Y + c$

Defining parameters†

A *parameter* is (formally) an operation on a super-population, mapping it to a "parameter space" Θ, such as $\mathbb{R}$, or $\mathbb{R}^p$, or $\{0, 1\}$.

The parameter *value* (typically denoted β or θ) is the result of this operation[5].

- "Inference" means making one or more conclusions about the parameter value
- These could be estimates, intervals, or binary (Yes/No) decisions
- "*Statistical* inference" means drawing conclusions **without** the full populations' data, i.e. in the face of uncertainty. Parameter values themselves are fixed unknowns; they are not "uncertain" or "random" in any stochastic sense.

In previous courses, parameters may have been defined as linear operations on the superpopulation. In 754, we will generalize the ~~idea~~.

[5] *The "true state of Nature" is a common expression for the same thing*

Defining parameters†

In this course, we will typically assume relevant parameters can be identified in this way. But in some real situations, one cannot identify θ, even with an infinite sample (e.g. mean height of women, when you only have data on men)

If your data do not permit useful inference, you could;

- Switch target parameters
- Extrapolate cautiously i.e. make assumptions
- Not do inference, but "hypothesis-generation"
- Give up

I will mainly disucss "sane" problems; this means ones we can reasonably address. Be aware **not** every problem is like this...

The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data
-John Tukey

Defining parameters†

Of course, infinite populations are an abstraction. But formally [6] statistical inference is [mostly] about parameter values determined from e.g.

- The heights of all men aged 50-100, this year and in all years, ever
- The heights of all possible men aged 50-100, in this and all possible universes
- The heights of all possible men aged 50-100 in Maryland, in this and all possible universes

Naturally, these abstract notions are not usually discussed in practice but thinking about $n = \infty$ will be helpful, when deciding exactly what parameters are of interest.

[6] When discussing [most] practical problems with your co-authors, it won't hurt to replace the infinite super-population with a vast substitute e.g. all men aged 50-100 in the US, or in developed countries

What is regression?†

In its most fundamental interpretation, regression estimates differences in outcome Y , between subjects whose X values differ in a specified manner.

We take differences in "Y" to mean differences in the expectation of Y , on some scale. For example, with binary X, you might be interested in;

$$\mathbb{E}_F[Y|X=1] - \mathbb{E}_F[Y|X=0]$$

or

$$log(\mathbb{E}_F[Y|X=1]/\mathbb{E}_F[Y|X=0])$$

or even

$$\exp\{\mathbb{E}_F[\log(Y)|X=1] - \mathbb{E}_F[\log(Y)|X=0]\}$$

Note that these are all different! As before, none of them is "right", "wrong", "uniformly best", or even "uniformly a great idea".

What is regression? : continuous X-values †

Q. How to concisely describe differences in Y over range of X?

The most commonly-used regression parameter is;

"The difference in Y per 1-unit difference in X"

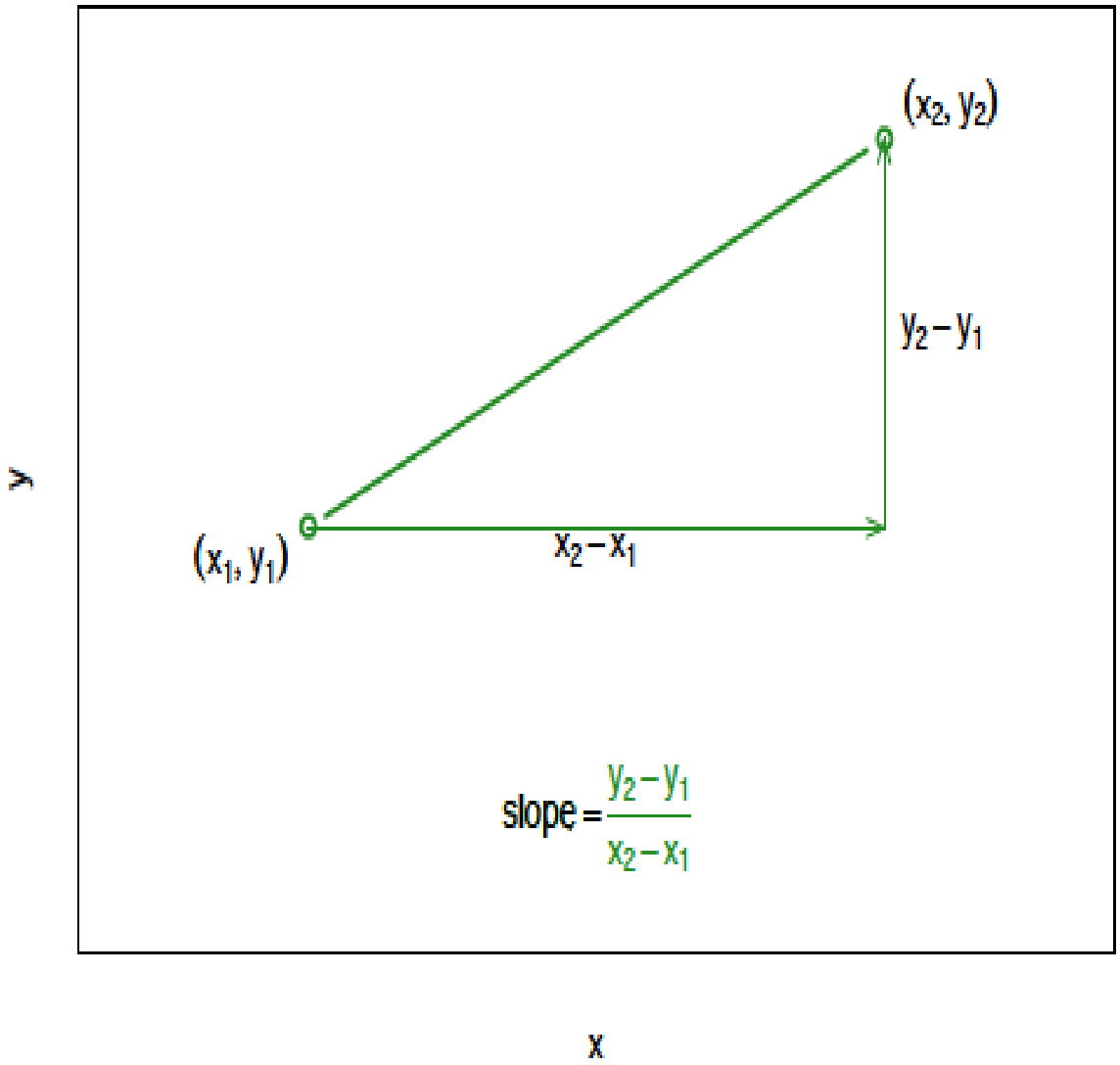
-which, most fundamentally, means:

- Take the difference in Y between two different X values *divided by* the difference in those X values
- Rinse and repeat, averaging this "slope" over all pairs of $\{Y, X_j\}, \{Y, X_k\}$.

(Other interpretations will be given later)

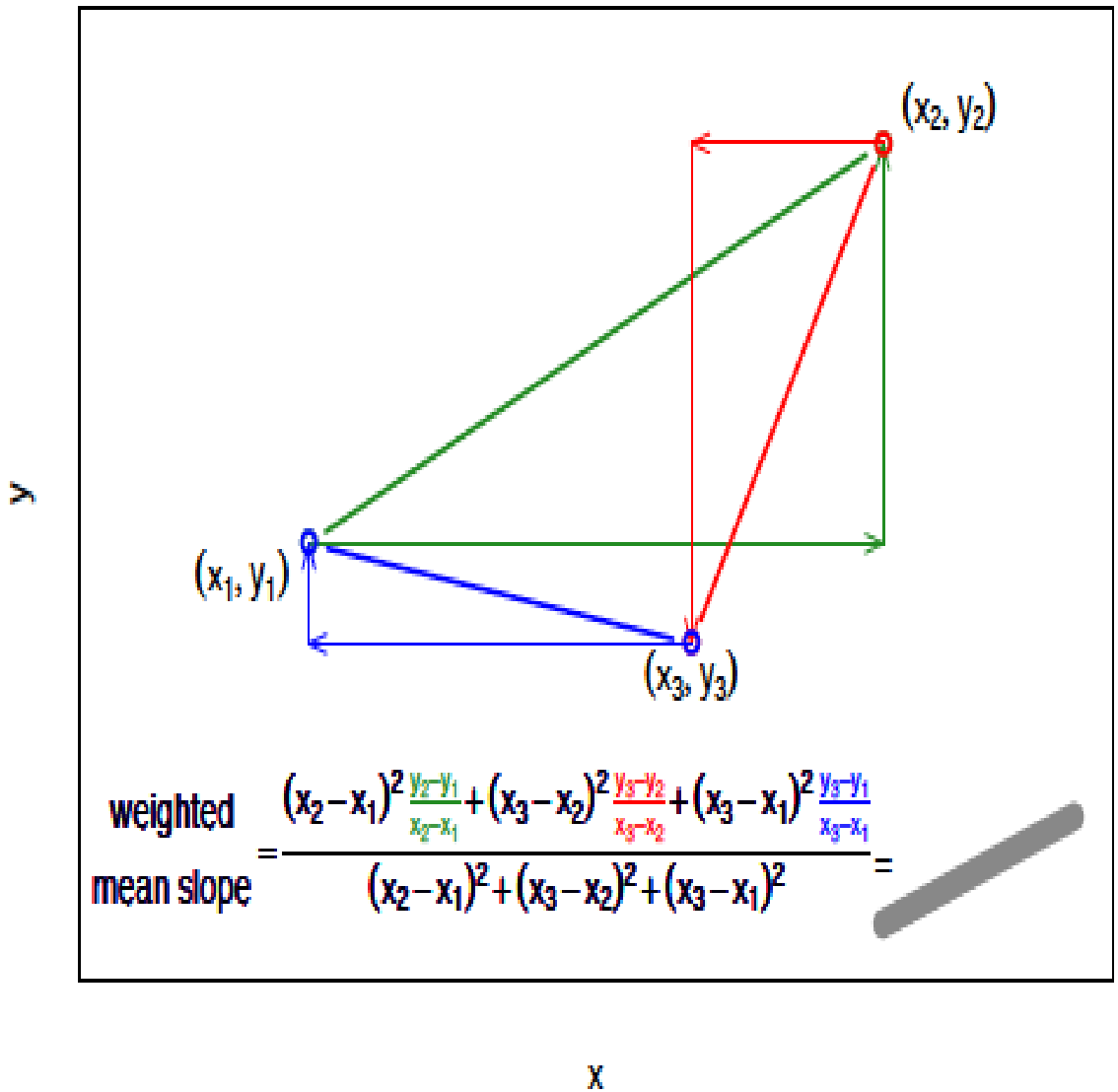
What is regression?: 2 X-values†

In a universe of only two points:



What is regression?: more X-values†

Default "averaging" uses weights $\propto (X_j - X_k)^2$:



What is regression?: many X-values†

Jacobi[7] showed there is a neater way to define the weighted mean slope parameter:

$$\beta_X = \frac{\mathrm{Cov}_X[X, Y]}{\mathrm{Var}_F[X]}$$

It can also be described as a (partial) solution to this system of equations:

$$\begin{aligned} \mathbb{E}_F[\beta_0 + X\beta_X] &= \mathbb{E}_F[Y] \\ \mathbb{E}_F[X(\beta_0 + X\beta_X)] &= \mathbb{E}_F[XY], \end{aligned}$$

where β_0 is a "nuisance" parameter; without further information, its value doesn't tell us anything about β_X. Please don't misinterpret the term "nuisance" to mean "totally useless" or "never of any interest".

[7] *... in 1841; the result is often overlooked*
Jacobi CGJ; De formatione et proprietatibus detrminantum.

What is regression?: many X-values†

How would we estimate β_X? We will assume (for now) that F denotes taking a simple random sample from the superpopulation. An "empirical" or "plug-in" estimate substitutes $\mathbb{E}_F$ with summation over the sample, hence:

$$\hat{\beta}_X(Y, X) = \frac{Cov(X, Y)}{Var(X)}$$

for sample covariance, variance. Equivalently, it solves

$$\begin{aligned} \sum_{i=1}^{n} \hat{\beta}_0 + X_i \hat{\beta}_X &= \sum_{i=1}^{n} Y_i \\ \sum_{i=1}^{n} X_i(\hat{\beta}_0 + X_i \hat{\beta}_X) &= \sum_{i=1}^{n} X_i Y_i \end{aligned}$$

-both forms should look familiar. Note you can express them (better) in matrix form:

$$X(Y - X\hat{\boldsymbol{\beta}}) = \mathbf{0}$$

where Y is $n \times 1$, and $X_{n\times 2}$ has row entries $\{1, X_i\}$.

What is regression?: no closed form†

The "difference" parameter need not be available in closed form. For example, for $Y > 0$ we may want to know;

$$\theta > 0 : \theta = \frac{\mathbb{E}_F[XY]}{\mathbb{E}_F[Y]} / \frac{\partial}{\partial \theta} \log \left(\mathbb{E}_F[\theta^X] \right)$$

- which tells you about multiplicative differences in Y, for different X values. Introducing the parameter θ_0, it can also be written

$$\begin{aligned} \mathbb{E}_F[Y - \theta_0 \theta^X] &= 0 \\ \mathbb{E}_F[X(Y - \theta_0 \theta^X)] &= 0 \end{aligned}$$

which can be combined in a vector equation

$$\mathbb{E}_F \left[\{1, X\}^T (Y - \theta_0 \theta^X) \right] = 0$$

What is regression?: no closed form†

An empirical estimate of θ is given by solving

$$\sum_{i=1}^{n}(Y_i - \hat{\theta}_0\hat{\theta}_i^X) = 0$$

$$\sum_{i=1}^{n} X_i(Y_i - \hat{\theta}_0\hat{\theta}_i^X) = 0$$

or writing $\boldsymbol{\beta} = \{\log\theta_0, \log\theta\}$ and using matrices:

$$X^T(Y - e^{X^T\hat{\boldsymbol{\beta}}}) = 0$$

where as before, $X_{n\times 2}$ has row entries $\{1, X_i\}$ and Y is $n \times 1$.

- There is no close form solution for $\hat{\beta}$ - but everything is perfectly well defined and "sane"
- In 2011 (not in 1959) numerical solutions are easily obtained if you know X, Y (Having real-valued parameters helps)
- Lack of closed form $\hat{\boldsymbol{\beta}}$ doesn't stop us working out/estimating frequency properties for $\hat{\boldsymbol{\beta}}$.

Defining parameters†

Some more complex parameters, defined via superpopulations;

1. The average $\Delta(Y)/\Delta(X)$, averaging pairs of observations - and weighting this average proportionally to $\Delta(X)^2$.
2. The least squares fit to the line $Y = g(X^T\beta)$.
3. The weighted least squares fit to the line $Y = g(X^T\beta)$, weighted by some $w(X^T\beta)$
4. As above, except we minimize by iteratively reweighted least squares, and not "proper" minimization (!)

(Throughout, I will assume that $\beta \in \mathbb{R}^p$)

Defining parameters†

Here are mathematical definitions:

1. $\boldsymbol{\beta} = \underset{\boldsymbol{\beta}'}{\text{argmin}} \mathbb{E}_F \left[(Y - X^T\boldsymbol{\beta})(Y - X^T\boldsymbol{\beta}') \right]$
   $= \mathbb{E}_F \left[XX^T \right]^{-1} \mathbb{E}_F[XY]$
2. $\boldsymbol{\beta} : \mathbb{E}_F \left[\frac{\partial g(X^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left(Y - g(X^T\boldsymbol{\beta}) \right] = \mathbf{0} \right.$
3. $\boldsymbol{\beta} : \mathbb{E}_F \left[\frac{\partial g(X^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} w(X^T\boldsymbol{\beta}) \left(Y - g(X^T\boldsymbol{\beta}) \right) \right] = \mathbf{0}$
4. $\boldsymbol{\beta} =$
   $\lim_{k \to \infty} \left\{ \boldsymbol{\beta}^{[k+1]} := \underset{\boldsymbol{\beta}'}{\text{argmin}} \mathbb{E}_F \left[w(X^T\boldsymbol{\beta}^{[k]})(Y_i - g(X^T\boldsymbol{\beta}'))^2 \right] \right\}$

... in all cases, F denotes simple random sampling of $\{Y, X\}$ from the superpopulation. Also, all equations are p-dimensonal.

Defining Parameters By Equations†

The general form of these equations is:

$$\mathbb{E}_F[\mathbf{G}(\beta, Y|X)] = \mathbf{0}$$

where $\mathbf{G}()$ maps to $\mathbb{R}^p$. Typically $\mathbf{G}()$ involves an expression in $Y - g(X^T\beta)$, somewhere.

Without any parametric assumptions, we are defining regression parameters β as quantities reflecting the difference in Y **associated with** some specific difference in X.

Formally we are defining β as a **functional** of F. For convenience, we assume that a unique root β exists; having multiple roots or no roots can happen - and theory exists to cope - but these are seldom a major problem in practice.

Link functions†

The 'link' function $g^{-1}()$ indicates how we are measuring differences in Y;

- Additive differences $\Leftrightarrow$ Identity link
- Multiplicative differences $\Leftrightarrow$ Log link

For odds ratios, the logistic link specifies:

$$g(X^T\beta) = \frac{\exp(X^T\beta)}{1 + \exp(X^T\beta)}$$

and is commonly used with binary Y.

The complementary log-log link specifies

$$g(X^T\beta) = \exp\left(-e^{X^T\beta}\right)$$

and is most-often used when Y is time to event.

Estimating Parameters†

Defining parameters is a first step; next we want to estimate these parameters.

As F provides data "rows" $\{Y_i, X_i\}$ as independent random samples, the expectations above are easily "mimicked"; for a sample of size n from F, an "empirical" (and generally **sane**) estimator $\hat{\boldsymbol{\beta}}$ can be defined as the solution to the 'estimating equation' (EE):

$$\sum_{i=1}^{n} \mathbf{G}(\hat{\beta}, Y_i, X_i) = \mathbf{0}$$

G is known as the "estimating function"; it is vector valued and maps to $\mathbb{R}^p$.

Solve the EE(s) gives p-dimensional $\hat{\boldsymbol{\beta}}$.

Estimating parameters†

Here are mathematical definitions:

1. $\hat{\boldsymbol{\beta}} = \underset{\boldsymbol{\beta}}{\text{argmin}}(Y - X^T\boldsymbol{\beta})(Y - X^T\boldsymbol{\beta})$
   $= (X^T X)^{-1} X^T Y$
2. $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} : \sum_{i=1}^{n} \frac{\partial g(X_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} \left(Y_i - g(X_i^T\boldsymbol{\beta})\right) = \mathbf{0}$
3. $\hat{\boldsymbol{\beta}} = \boldsymbol{\beta} : \sum_{i=1}^{n} \frac{\partial g(X_i^T\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} w(X_i^T\boldsymbol{\beta}) \left(Y_i - g(X_i^T\boldsymbol{\beta})\right) = \mathbf{0}$
4. $\boldsymbol{\beta} =$
   $\lim_{k\to\infty} \left\{ \boldsymbol{\beta}^{[k+1]} := \underset{\boldsymbol{\beta}'}{\text{argmin}} \sum_i w(X_i^T\boldsymbol{\beta}^{[k]})(Y_i - g(X_i^T\boldsymbol{\beta}'))^2 \right\}$

At least for 1-3, we are just replacing F by the empirical distribution function defined by our data. Use 4. to justify parameters in terms of Iteratively Re-weighted Least Squares (IWLS) ... if you must.

Properties of the estimates from the CLT †

For general $\boldsymbol{\theta}$ satisfying $\mathbb{E}_F[\mathbf{G}(\boldsymbol{\theta}, Y, X)] = \mathbf{0}$, we use estimating equations:

$$\sum_{i=1}^{n} \mathbf{G}(\hat{\beta}, Y_i, X_i) = \mathbf{0}$$

Many similar size "contributions" are being added, the Central Limit Theorem is therefore useful for derivingthe frequentist properties of estimating function $\mathbf{G}(\cdot, \cdot, \cdot)$ These properties can be *transferred* to the resultant estimator $\hat{\theta}$, allowing us to specify:

- Large sample limiting value of $\hat{\theta}$
- Large sample variance of $\hat{\theta}$
- Large sample distribution of $\hat{\theta}$

These can be used to give (valid) large-sample confidence intervals, whatever the true-but-unknown F, or $\boldsymbol{\theta}(F)$.

Standard error estimates: theory †

Suppose that, based on a sample of size n, $\hat{\boldsymbol{\theta}}_n \in \mathbb{R}^p$ is a solution to the estimating equation $\sum_{i=1}^n \mathbf{G}(\boldsymbol{\theta}, Y_i, X_i) = \mathbf{0}$. Under mild regularity conditions, $\hat{\boldsymbol{\theta}}_n \to_P \boldsymbol{\theta}$ - so $\hat{\theta}_n$ is a consistent estimate of $\boldsymbol{\theta}$. Furthermore:

$$\sqrt{n}(\hat{\boldsymbol{\theta}}_n - \boldsymbol{\theta}) \to_D N_p(\mathbf{0}, \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{T-1})$$

where

$$\begin{aligned} \mathbf{A} &= \mathbf{A}(\boldsymbol{\theta}) = \mathbb{E}_F\left[\frac{\partial}{\partial \boldsymbol{\theta}}\mathbf{G}(\boldsymbol{\theta}, Y, X)\right] \\ \mathbf{B} &= \mathbf{B}(\boldsymbol{\theta}) = \mathbb{E}_F\left[\mathbf{G}(\boldsymbol{\theta}, Y, X)\mathbf{G}(\boldsymbol{\theta}, Y, X)^T\right] = \mathrm{Cov}_F[\mathbf{G}(\boldsymbol{\theta}, Y, X)] \end{aligned}$$

This means $\hat{\boldsymbol{\theta}}$ is asymptotically Normal, around the "right" mean, with a variance that shrinks with n^{-1}.

Standard error estimates: theory†

- $\mathrm{Var}_F[\hat{\theta}_n] \approx \mathbf{A}^{-1}\mathbf{B}\mathbf{A}^{T-1}/n$ is known as the "sandwich formula". $\mathbf{A}^{-1}$ is informally known as the "bread", and $\mathbf{B}$ is the "meat".
- "Mild" really is "mild"; a few moment conditions will typically suffice
- The CLT is your friend! For many problems, the approximations are **very good** for n in the hundreds - but for $n < 10$ don't expect miracles.
- The asymptotics of location/spread can "kick in" at different rates. For "hard" problems Normality may be a poor approximation to the behavior of $\hat{\theta}$ unless n is **vast**.

Standard error estimates: the sandwich†

- The previous result is very important; it tells us that, with large n, the distribution of $\hat{\theta}^*$ will be centered around θ, the value we want to know, and "spread" in a manner we understand well.
- However, the **A** and **B** terms in the result are expectations involving expressions in (fixed-but-unknown) θ, over (also unknown) F. Without *very* strict, "pretty" restrictions on F, exact evaluation of these expectations is hopeless.
- Nevertheless, in large samples, **A** and **B** can be (very) well estimated; we plug in $\hat{\theta}_n$ for the true parameter value θ, and use averaging over our dataset to substitute for expectation over F.

Standard error estimates: the sandwich†

If we plug-in empirical estimates of $\boldsymbol{\theta}$ and F, i.e.,

$$\begin{aligned}\hat{\mathbf{A}} &= \frac{1}{n}\sum_{i=1}^{n}\frac{\partial}{\partial\boldsymbol{\theta}}\mathbf{G}(\hat{\boldsymbol{\theta}}_n, Y_i, X_i)\\ \hat{\mathbf{B}} &= \frac{1}{n}\sum_{i=1}^{n}\mathbf{G}(\hat{\boldsymbol{\theta}}_n, Y_i, X_i)\mathbf{G}(\hat{\boldsymbol{\theta}}_n, Y_i, X_i)^T\end{aligned}$$

then (by a law of large numbers) $\hat{\mathbf{A}} \to_P \mathbf{A}$ and $\hat{\mathbf{B}} \to \mathbf{B}$, so

$$\widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_n) = \frac{1}{n}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\hat{\mathbf{A}}^{T-1}$$

is a **consistent** estimator of the variance of $\hat{\theta}_n(Y)$. Intervals based on $\hat{\theta}_n \to_D N_p(\boldsymbol{\theta}, \widehat{\mathrm{Var}}(\hat{\boldsymbol{\theta}}_n))$ have the correct coverage, asymptotically.

This is known as the **sandwich covariance estimate** due to Huber (1967, Proc 5th Berk Sym) - and Eicker, and White. Hansen (1982, Econometrika) proposed the general form.

Standard error estimates: the sandwich†

- Also known as the "robust" estimate of (co)variance, used in "robust standard errors" and "robust intervals"
- As it can behave badly in some (non-asymptotic) situations, "model-agnostic" is better; we're using **no** parametric ideas
- Also known as a "heteroskedasticity-consistent" estimate. This name:
  - badly understates the utility; we specified almost nothing about F - why worry only about the variance?
  - regularly defeats seminar speakers
- EE and the sandwich are known as the Generalized Method of Moments in econometrics where they are common.But they were largely unknown to statisticians before Royall (1986, Intl Stat Rev)