review session

gov 2000

Overview

- **Random Variables and Probability**
- Univariate Statistics
- Bivariate Statistics
- Multivariate Statistics
- Causal Inference

Probability

Why is probability important?

- Probability involves reasoning about samples given the truth. (Ex. You know the prob of a coin flip is $\frac{1}{2}$. You then ask how many H's you expect to get if you flip the coin six times.)
- Inference is about reasoning about the truth given a sample. (Ex. You've fliped a coin six times and gotten five heads. Could you reasonably expect to have gotten that with a fair coin?)
- Probability and inference are the two sides of the "same coin."

Random variables

(From Wikipedia) A random variable can be thought of as an unknown value that may change every time it is inspected. Thus, a random variable can be thought of as a function mapping the sample space of a random process to the real numbers. A few examples will highlight this:

- A coin toss with H or T being the two equally likely events;
- The toss of a die with the numbers 1-6 being the equally likely outcomes;
- A spinning wheel that can choose a real number from the interval [0, 2π), with all values being equally likely.

Random variables – ctd

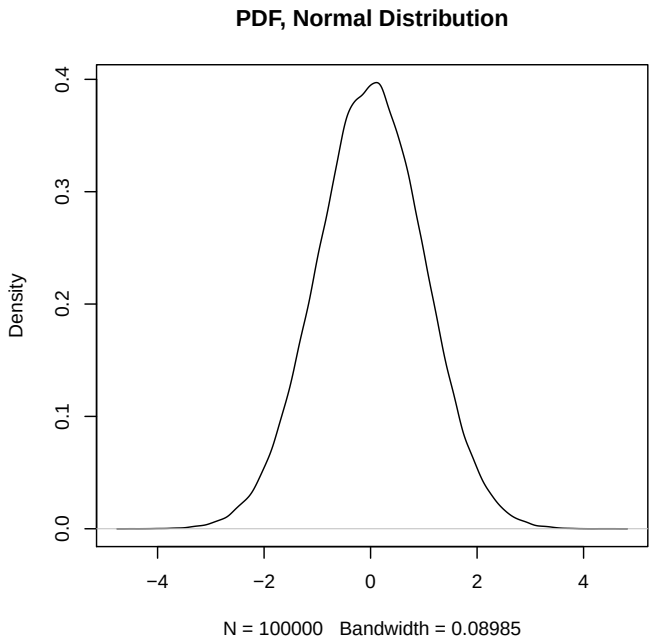
- For **discrete distributions**, the random variable X takes on a finite, or a countably infinite number of values. (Ex. The number of H's in six coin flips.)
- For **continous distributions**, the random variable X takes on continuous or infinte values. (Ex. The distribution of weights of Gov2000 students.)

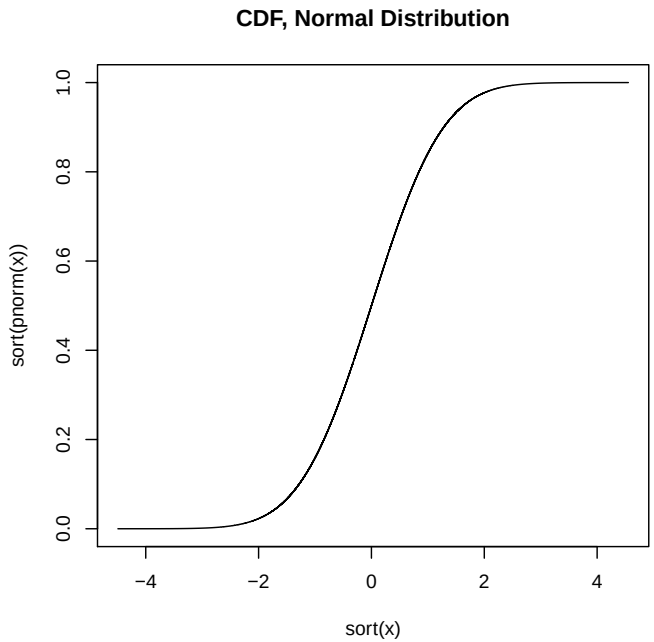
Random variables – ctd

- A probability mass/density function (PMF/PDF) and a cumulative mass/density distribution function (CMF/CDF) are two common ways to define the distribution for a discrete RV.
- The PMF/PDF, $f(x)$, is a function that describes the relative likelihood for this random variable to occur at a given point in the observation space.
- The CDF/CDF, $F(x)$, gives the probability that the random variable X takes on a value less than or equal to x.

Graphical representation

It's sometimes easier to understand the PDF and CDF graphically:





Overview

- Random Variables and Probability
- **Univariate Statistics**
- Bivariate Statistics
- Multivariate Statistics
- Causal Inference

Univariate statistics

We started off talking about inference by covering univariate statistics – when you only have one variable of interest. Here are some examples:

- You want to estimate level for support for one candidate in the general population;
- Your want to estimate the subprime lending rate in a depressed county in Florida;
- You want to estimate the mean democracy score for countries in South America.

This was material covered up through the midterm.

Terminology

At this point, we introduced

- **Parameters**: characteristics of the population distribution (e.g. the mean), often denoted with a greek letter (μ, σ^2).
- **Estimators:** Random quantities, written as X or Y.
- **Estimates:** Realized values of an estimator; hence they are not random (e.g. $\bar{x}$).

Sampling distributions

A key concept is that of a sampling distribution. Here's how it works.

- We have a large population ("The Truth").
- We take a sample from the population (usually, as researchers, this is what we see – a sample).
- We calculate our estimate (using our estimator) from this sample.
- And then put the sample back, draw another, and calculate our estimate again.
- We repeat.

Be sure you understand this concept, as it drives our thinking about confidence intervals and hypothesis testing.

Estimating population mean

We usually want to estimate the population mean. Let's use as our estimator the mean of the sample, $\bar{x}$.

- Our point estimate will be $\bar{x}$.
- Under repeated sampling, $\bar{x}$ will be normally distributed Why? Central Limit Theorem!
- Because $\bar{x}$ is normally distributed, we can construct confidence intervals and hypothesis tests.

Takeaway: The Central Limit is amazing and makes many inferences possible.

Confidence intervals

To construct confidence intervals for the true population parameter:

- For a large sample, use the we construct the $(1 - \alpha)$ confidence interval with $\bar{x} \pm z_{\alpha/2} SE$.
- For a small sample $(n < 32)$, we use $\bar{x} \pm t_{\alpha/2} SE$.
- Remember the correct interpretation for a confidence interval – if you repeatedly drew samples, and for each sample constructed a confidence interval, then $1 - \alpha$ percent of the intervals would contain the true population parameter.

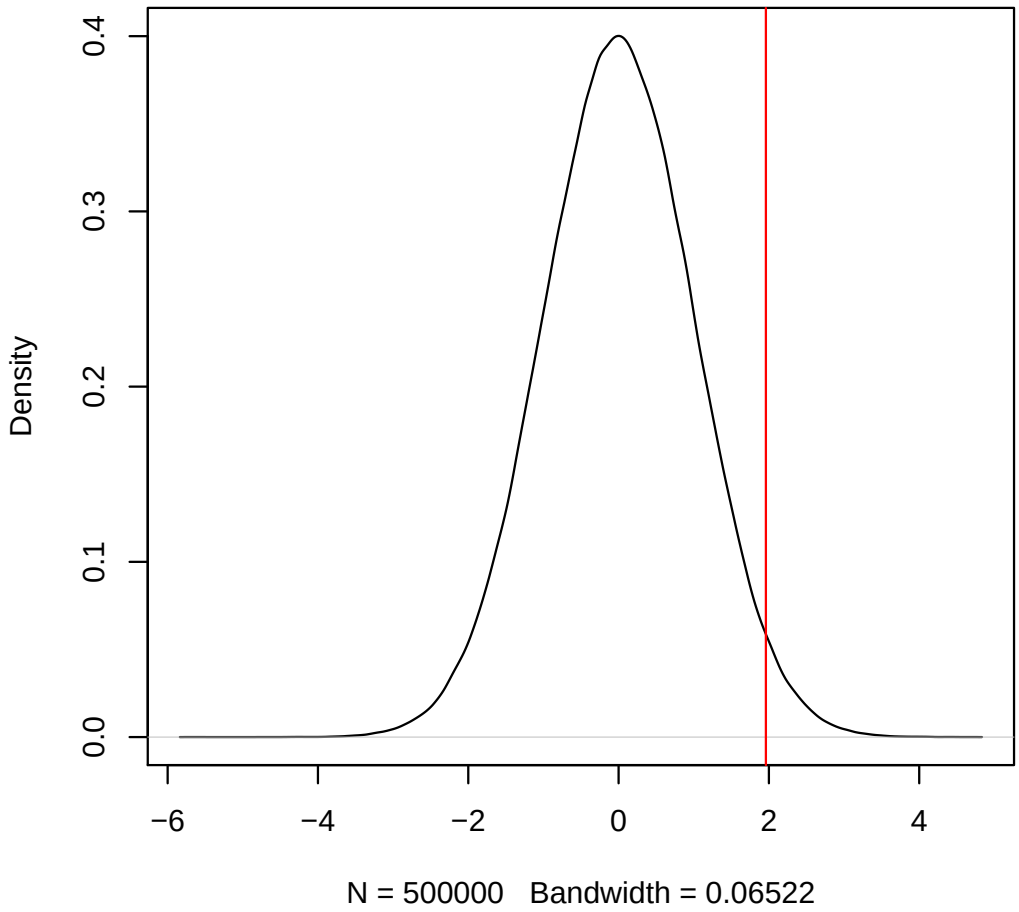
Hypothesis testing

The CLT also allows us to construct hypothesis tests. Most frequently, we want to test the null that the parameter is zero. Here's how it works:

- We assume that the observed estimate comes from the null distribution. We then transform the null (which comes from a normal) into a standard normal by substracting the mean and dividing by the standard error. The mean under the null is zero.
- This gives you your test statistic, $\frac{\bar{x}-0}{SE}$.
- Then you compare this to a standard normal to see how unusual this observation would be. If it's pretty unlikely to happen, then we reject the null.
- Note that the p-value just quantifies this – it's the probability of obtaining a test statistic at least as extreme as the one that was actually observed

Hypothesis testing graphically

It's again sometimes easier to see this graphically:



Overview

- Random Variables and Probability
- Univariate Statistics
- **Bivariate Statistics**
- Multivariate Statistics
- Causal Inference

Bivariate regression

We went from being interested in just one variable to being interested in two variables and the relationship between them.

- Univariate examples: Estimating the support for a candidate in a population, estimating the true subprime mortgage lending rate
- Bivariate examples: The relationship between the unemployment rate and suicide, the relationship between perceptions of CEO pay and ideal CEO pay.

Note: The term regression, which we used moving forward from this point, can be used to describe the relationship between any independent and dependent variables.

Introducing least squares

We discussed various techniques for summarizing the relationship between two variables (LOESS), etc., but we decided that ordinary least squares regression would be the best.

- How do we derive OLS? We plot the data, then fit the line that minizes the vertical distance between the line and the actual observations.
- Check out the calculus in Adam's slides. The point is that we can get estimates for the line's intercept, β_0 and slope, β_1. We are especially interested in the slope, as it summarizes nicely the relationship between X and Y.

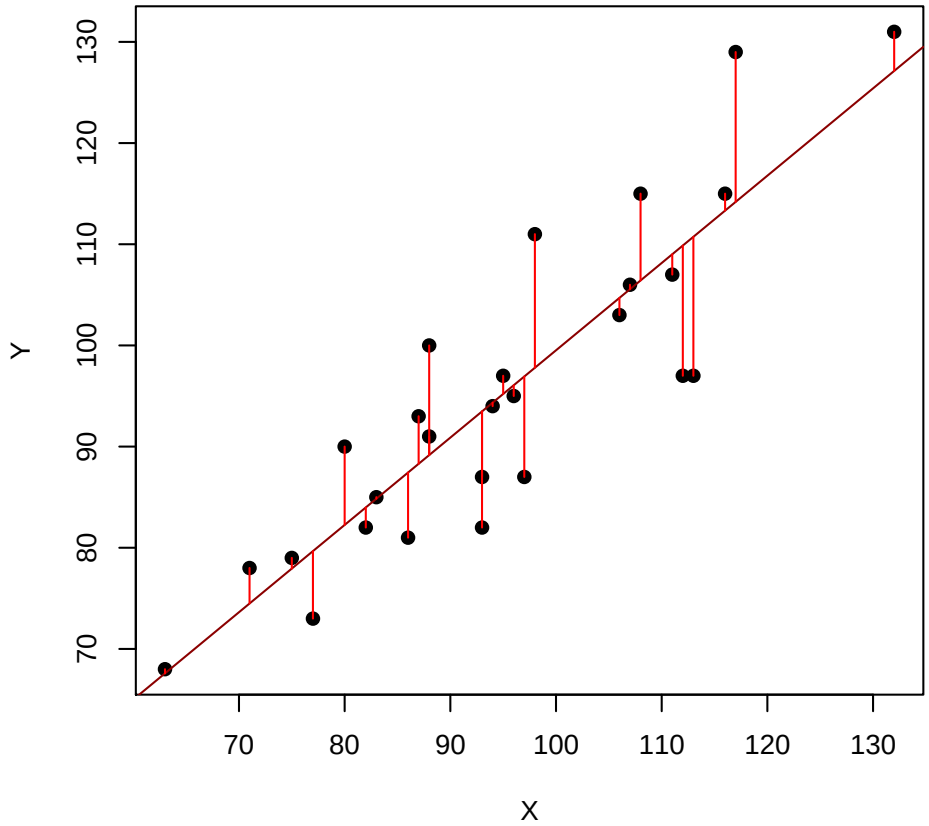
Least squares – ctd.

More specifically,

- We end up with a line that looks like this: $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X_i$.
- $\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$
- $\hat{\beta}_1 = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{x_i - \bar{x}}$
- You can also calculate r^2, how much variation in Y is accounted for by X.

A graphical representation of OLS

It's sometimes easier to understand this graphically:



OLS repeated sampling

Again, think about this as being in the context of sampling. We draw one sample from the true population, calculate OLS estimates, put the same back, draw another sample, calculate OLS estimates again, etc.

- In repeated sampling, $\hat{\beta}_1$ and $\hat{\beta}_0$ will be normally distributed due to the Central Limit Theorem.
- More specifically, $\hat{\beta}_1 \sim N(\beta_1, \frac{\sigma^2}{\sum(x_i - \bar{x})})$.
- This means that we can use the same techniques as before to express uncertainty around our $\hat{\beta}_1$ and $\hat{\beta}_0$ estimates.

Confidence intervals for OLS estimates

Because $\hat{\beta}_1$ and $\hat{\beta}_0$ are normally distributed, we can construct confidence intervals.

- For the slope: $\hat{\beta}_1 \pm t_{\alpha/2}\hat{SE}[\hat{\beta}_1]$
- For the intercept: $\hat{\beta}_0 \pm t_{\alpha/2}\hat{SE}[\hat{\beta}_0]$
- The interpretation is the same as before – with repeated calculations, we'd expect that $1 - \alpha$ percent of the confidence intervals we create would contain the true values of β_1 and β_0.

Hypothesis testing for OLS estimates

The most tested null is that the slope is equal to zero; that is, that there is *no relationship* between the independent and dependent variables ($H_0 : \beta_1 = 0$).

- As before, we know the null is normal. We want to get everything to a standard ($N(0, 1)$) normal, so we standardize everything, including our coefficient. We use the same formula, $\frac{\hat{\beta}_1 - 0}{\hat{SE}[\hat{\beta}_1]}$
- We then compare this to a standard normal to see how likely (or unlikely) the coefficient would be under the null being true.

Overview

- Random Variables and Probability
- Univariate Statistics
- Bivariate Statistics
- **Multivariate Statistics**
- Causal Inference

Multivariate regression

We went from being interested in just one variable to being interested in two variables to now being interested in many independent variables (the covariates, or X variables) and one dependent variable (Y).

- Univariate examples: Estimating the support for a candidate in a population.
- Bivariate examples: Relationship between the unemployment rate and suicide.
- Multivariate examples: Relationship between British colonial history, Islam, and democracy scores; relationship between unemployment, inflation, and candidate support.

Note: Most social science research is multivariate.

Multivariate regression – ctd.

Everything basically works the same way, except that now we minimize the square of the residual with respect to the regression plane (not line).

- The regression line looks like $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_{1i} + \hat{\beta}_2 X_{2i} + ... + \hat{\beta}_k X_{ki}$.
- Also, all of the $\hat{\beta}$'s are normally distributed, so we can make the same kinds of infereces as before. (Just note that the standard error is now adjusted by the variance inflation factor – see lecture notes!)
- As before, we can calculate r^2, but we might want to use an ajudsted r^2 to copensate downward for the number of additional covariates.

Interaction terms

Something that comes up a lot in political science is the use of interaction terms.

- Interactions should be used when you think the "effect" of one variable increases or decreases along with another variable.
- $\hat{y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i + \hat{\beta}_2 X_i + \hat{\beta}_3 (X_i * Z_i)$
- $\hat{\beta}_1$ now represents the coefficient on X when $Z = 0$, and $\hat{\beta}_2$ is the coefficient on Z when $X = 0$.
- When you are unusure of how to interpret a regression output involving an interaction term, write out the regression equation and use algebra to help you.
- Never fail to include lower order terms unless you think carefully about the substance of the constraints you are placing on your model.

Matrix algebra

A quick note about matrix notation:

- We use matrix notation ($\mathbf{y}$, $\mathbf{x}$, β) because it gets really cumbersome to write out $y_1, y_2, ..., y_n$.
- Using matrix notation we get the same OLS estimator as before, only now it's in matrix form: $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'\mathbf{y}$
- It also makes it a lot easier to do things like F tests.

F test

- The F test makes it handy to test multiple hypothesis – for example, testing a null that a subset of your coefficients equal zero, or if a subset of your coefficients equal each other.
- To see the specifics of how to do an F test, go through Adam's notes.
- You can also use the `linear.hypothesis` function in the `car` library in `R`.

Diagnostics

- Diagnostics are an important final step for both bivariate and multivariate regressions.
- After you're done with your analysis, check for possible leverage points, influence points, and outliers.
- Also check that the modeling assumptions hold.
- In particular, check possible non-constant error variance. If you have evidence of non-constant error variance, then you might want to use Huber-White standard errors.

Overview

- Random Variables and Probability
- Univariate Statistics
- Bivariate Statistics
- Multivariate Statistics
- **Causal Inference**

Causal inference

So far, this review session has dealt exclusively with predictive inferences, not causal ones. Making causal inferences is much harder. Why?

- Making causal inferences involves many more assumptions.
- You have to think carefully about what your "treatment" is and what your treatment and control groups are.
- You have to think carefully about the moment of treatment and which of your variables and pre-treatment and which are post-treatment.
- You have to consider whether other factors not included in your model could be driving the results.

Key assumptions

- Ignorability. You must be able to assume that there are no confounding variables that could affect both the probability of treatment and the outcome variable.
- SUTVA (Stable Unit Treatment Variance Assumption). Make sure that your treatment isn't seeping from one subject to another.
- Post-treatment issues. Also be wary of introducing into your model variables that take place after the treatment has been administered.

Causal inference – bottom line

- If you have a randomized experiment then you can assume that there are no confounders.
- Political scientists often look for "natural" experiments – like the one in Andy's paper. These can be really useful, but be sure to justify the rasons why you think the treatment is still random.
- If you have an observational study, you'll want to control for all the confounders that you possibly can. Making causal inferences becomes a lot more difficult.
- In both contexts, you'll want to think about possible SUTVA violations and be careful about not introducing post-treatment bias.
- In general, be very careful about making causal claims and consider other techniques – matching, regression discontinuity, experimentation.

Conclusion

- Random Variables and Probability
- Univariate Statistics
- Bivariate Statistics
- Multivariate Statistics
- Causal Inference

Immediately following this course

- Gov 2001 – will cover dichotomous dep variables, maximum likelihood, as well as key topics (missing data, matching). You'll go through a replication of a published paper from beginning to end.

Completing the 2000/2001 sequence will put you at the forefront of mainstream political scientists.

Down the road

- Stat 110/210 – probability theory (essential for understanding the foundations of likelihood and bayesian; no computation)
- Stat 111/211 – inference (a lot of proofs involving what we learned in this class and will learn in Gov 2001; no computation)
- Gov 2003 – Bayesian (when offered; lots of computation)

Completing this sequence in addition to the 2000/2001 sequence will put you at the forefront of quant-oriented political scientists.

Other courses to take

- Stat 245 – statistics and litigation
- Econ 1128 – Don Rubin's causal inference class
- Gov 2010 – research methodology
- Gov 2002 – topics (when offered, usually more advanced causal inference, textual analysis)
- Stat 139/149 – generalized linear models (similar to what's covered in 2001, but more focus on the math, less on computation)
- Stat 220 – Bayesian
- Stat 221 – Bayesian computing

Taking all these courses would basically make you a methodologist.