DL 902 — Research Methods II: Statistics and Causal Inference

The quantitative half of the methods core. DL 901 taught you to design a study that can answer the question. This course teaches you to analyze the evidence the design collects — and to know when the evidence is not strong enough to say what you hoped.

Level: doctoral · **Prerequisites:** DL 901 (or equivalent graduate research-design course) · **Companion:** DL 903 (empirical study-design seminar) · **Credits:** 3 · **Format:** one 3-hour seminar per week, one term.

This course is graduate-level statistics and econometrics aimed squarely at the field's data: **binary outcomes** (won / lost), **continuous outcomes** (award value, win rate), **hierarchical data** (pursuits nested inside firms inside agencies), and the **inference hazards** that make this field's causal claims hard. It is taught at concept altitude — you learn the statistical ideas, what they are for, what they assume, and how to read them honestly. No lesson requires operating any software; the methods are the subject, and the tools are swappable (doctrine/08).

Course description

The field's core empirical act is comparing a **prediction** to an **outcome** — someone estimated a probability of winning, and the market produced an award. That simple act is the seed of almost every quantitative study in this discipline, and it hides real statistical depth:

- What does it *mean* for a probability estimate to be good — and how do you measure that?
- How do you model a binary outcome (won/lost) without fooling yourself?
- How do you attribute an outcome to a factor (certification, incumbency, price position) when firms and pursuits differ in a thousand unmeasured ways?
- How do you keep a finding honest when the data are small, clustered, and biased by who chose to pursue?

This course builds the quantitative toolkit for those questions. It is organized in four movements:

1. **Statistical foundations** — probability, distributions, estimation, testing, and power (a fast, graduate-level consolidation, not an introduction).
2. **Modeling outcomes** — regression, logistic regression, and the special craft of modeling win/loss data.
3. **Calibration and prediction quality** — the measurement of whether a probability estimate is any good: reliability, Brier score, expected calibration error, and population stability.

4. **Causal inference** — the potential-outcomes framework and the design-based tools (difference-in-differences, interrupted time series, regression discontinuity, instrumental variables) that let a researcher say "X *causes* Y" — or honestly say they cannot.

Learning outcomes

By the end of this course, a student can:

1. **Fit and interpret** a regression model for a continuous or binary outcome, and state the model's assumptions and limits.
2. **Measure prediction quality** with Brier score, log-loss, expected calibration error, and reliability diagrams — and interpret what each says about a probability estimator.
3. **Apply the causal-inference toolkit** to a real design question — choosing and justifying a difference-in-differences, interrupted-time-series, regression-discontinuity, or instrumental-variable approach — and state exactly what the approach assumes.
4. **Diagnose the field's bias traps** — selection, survivorship, confounding, clustering, and small-sample instability — in a given analysis, and correct for them or bound the claims.
5. **Preregister** an analysis plan and **write a methods section** that a replication-minded reader could execute.

Movement 1 — Statistical foundations (fast, graduate-level)

Probability and the idea of a model

A model is a simplified story about how the world generates the numbers you see. The most important habit this course installs is **always writing the model down** before touching data: "awarded ~ Bernoulli(p), where p depends on certification status, firm size, and NAICS." The model is the claim; the statistics test the claim.

We consolidate the foundations a graduate researcher must command cold:

- **Distributions** — the binomial/Bernoulli for won/lost, the lognormal for award values (award dollars are right-skewed; a $50M award is not "five times as likely" as a $10M one), the normal as the workhorse for test statistics.
- **Estimation** — estimators as rules, bias vs. variance, the standard error as the measurement of sampling luck.
- **Hypothesis testing and confidence intervals** — the logic of the null, what a p-value does and does not say (it is not the probability the null is true), and why confidence intervals carry more information than p-values.

- **Effect sizes** — a "statistically significant" 2-percentage-point win-rate difference may be trivial; the *size* of the effect is what matters for practice. Report the size, not just the star.
- **Power and sample size** — the probability that a study with N observations detects a true effect. The field's outcome data are often sparse, so power analysis at the design stage is not optional; it is what keeps a dissertation from dying quietly in the analysis phase.

Seminar activity. Given three published-style findings ("HUBZone firms won 58% of HUBZone set-asides, $p < 0.05$"), teams produce the sentence each finding *actually* supports, stripped of the overclaim. The exercise builds the habit of reading statistics as precisely as the discipline reads solicitation language.

Reading. doctrine/05 (the score's arithmetic as the field's native statistical object).

Deliverable. A one-page foundations memo: for the student's own outcome variable, state the distribution family, the estimator, and a power calculation for the smallest effect they would care to detect.

Movement 2 — Modeling outcomes

Regression

The workhorse. **Linear regression** models a continuous outcome as a weighted sum of predictors; the coefficients are the estimated associations, and the standard errors measure their uncertainty. The assumptions to interrogate: linearity, independence, constant variance, and no severe collinearity. In this field the biggest assumption failures are almost never the textbook ones — they are **confounding** (a predictor is correlated with an unmeasured cause of the outcome) and **clustering** (pursuits of the same firm are not independent).

Logistic regression for won/lost

Because the field's headline outcome is binary — won or lost — the logistic model is the natural instrument. It models the log-odds of winning as a weighted sum of predictors, and it returns a predicted probability in [0,1]. Every concept a practitioner already knows maps onto it:

- the predicted probability is a *pWin estimate*;
- the predictors are the *factors*;
- the coefficients are the *factor weights* that the field's scoring composites argue about by hand.

That mapping is the bridge between the doctrine's arithmetic and the research statistics: **a factor-weight validation study is a logistic regression that asks which predictors actually move the probability of winning, and by how much.** DL 903 builds the full study around this.

Modeling the messy reality

- **Clustered / panel data.** Pursuits are nested in firms; firms are nested in agencies' markets. Ignoring the nesting inflates significance. The fix — clustered standard errors, or a multilevel model with firm and agency random effects — is a required reflex for this field's data.
- **Award value as an outcome.** Log-transforming dollar outcomes; modeling win probability and award value *jointly* (the score is pWin × value, so a complete study often models both).
- **Missing and censored outcomes.** Pursuits that have not resolved yet; opportunities declined before scoring; withdrawn bids. Each is a data-availability decision with consequences (see the survivorship hazard, Movement 4).

Seminar activity. Teams receive a small real award dataset (public, anonymized) and a research question; they specify the logistic model, name the predictors, and — before running anything — list the three assumptions most likely to fail and what they would do about each.

Reading. doctrine/05; the glossary entries for *pWin, factor, threshold.*

Deliverable. A model-specification memo: the student's outcome, the candidate predictors (mapped to the field's factors), the model family, the clustering structure, and the stated assumptions.

Movement 3 — Calibration and prediction quality

This movement is the field's statistical signature. Capture is in the business of **probability estimates** — pWin — and probability estimates have a specific virtue that ordinary predictions do not: they can be *calibrated.* A well-calibrated 0.7 means the event happens 70% of the time.

Reliability and the reliability diagram

Bin the predicted probabilities (0.0–0.1, 0.1–0.2, . . .) and plot, in each bin, the *observed* win rate against the *predicted* center. A perfectly calibrated estimator sits on the diagonal. Deviations from the diagonal are calibration error, and where they occur tells you who the estimator is over- or under-confident about.

Brier score and log-loss

Two summary measures of a probability estimator's quality:

- **Brier score** — the mean squared difference between the predicted probability and the binary outcome, averaged over all predictions. Lower is better; a perfectly calibrated and perfectly discriminating estimator scores near

0; a coin flip scores 0.25. It rewards both *calibration* and *discrimination* (separating winners from losers).

- **Log-loss** — the negative log-likelihood of the observed outcomes under the predictions, penalizing confident errors harshly. A confident wrong prediction is worse than a hedged one.

Expected Calibration Error (ECE)

A sample-weighted average of the reliability-diagram deviations: how far, on average, the observed win rate is from the prediction, weighted by how many predictions fell in each band. It summarizes the calibration gap in one number, and it is the natural headline statistic for "is our pWin honest?"

Population Stability Index (PSI)

A drift statistic that compares the *distribution* of predictions in one period against a baseline. It does not measure calibration against outcomes — it measures whether the estimator is being used on a population that looks like the one it was built for. A large PSI says "the mix of what we are scoring has changed," which matters even before outcomes arrive. The standard reading: below 0.10 stable, 0.10–0.25 moderate shift, above 0.25 significant shift.

Calibration as a research subject

These are not just tools; they are the concepts behind a whole research thread. The discipline's machinery already encodes this doctrine: pWin estimates are meant to be calibrated to observed award rates, drift is monitored in bands (watch / warn / critical), and recalibration is done with a **monotone remap** that preserves ordering (a higher raw estimate must never map to a lower calibrated one — the pool-adjacent-violators / isotonic idea). For the doctoral student, the research questions write themselves: *Is the composite calibrated? In which bands is it over- or under-confident? Does calibration drift over time or across funder families?* Those questions are the core of the empirical-design seminar (DL 903).

Seminar activity. Given a small table of (predicted pWin, won/lost) pairs, teams hand-compute a Brier score and an ECE for a two-band reliability split, and interpret which band is the problem. The hand-computation installs the concept so no software is ever a black box.

Reading. doctrine/05; the concept-to-system map's scoring row (the composite and its calibration block) read as *concepts*.

Deliverable. A calibration memo: for the student's own study, state the prediction-quality measure they will report, the reliability-bin resolution, and the smallest calibration error they would consider a finding.

Movement 4 — Causal inference and the field's bias traps

The potential-outcomes framework

The clean way to state a causal question: each unit (firm, pursuit) has a *potential outcome* under treatment and a *potential outcome* under control; the causal effect is the difference. You observe only one. All of causal inference is the art of constructing a credible stand-in for the outcome you did not observe. This framing makes the field's hard problem obvious: **you never observe the outcome of a pursuit the organization did not pursue.**

The bias traps specific to this field

1. **Selection / survivorship.** Only pursued bids resolve. Any study using resolved outcomes is studying a sample chosen by the very decisions under study. *Mitigations:* model the selection where possible; stratify by funder family; treat withdrawal as a non-win by stated policy; and — always — bound the claims to pursued bids.
2. **Confounding.** Certified firms differ from non-certified firms in many ways beyond certification. *Mitigations:* adjust for measured confounders; exploit a design (below) that neutralizes unmeasured ones.
3. **Endogeneity of the treatment.** The firms that adopt color-team review may be the firms that were going to win anyway. The treatment is chosen, not assigned.
4. **Clustering and tiny-N instability.** A few firms' outcomes dominate a sample; a handful of extreme observations can flip a calibration curve. *Mitigations:* clustered inference; observation floors; pooling small cells.
5. **Feedback and oscillation.** Closed-loop "learning" systems that recalibrate on noisy windows can chase noise. *Mitigations:* bounded adjustments per cycle, minimum observation floors, and validation of the updated rule against the prior before promotion — the same discipline the field's machinery applies to its own recalibration.

The design-based toolkit

When a clean experiment is impossible, the researcher uses natural structure:

- **Difference-in-differences (DiD).** Compare the change in outcomes for a treated group before vs. after a treatment, against the change for a control group over the same period. *Example:* win rates of firms in the two years before vs. after obtaining HUBZone certification, compared to firms that never certified. Assumes parallel trends (the control group's counterfactual path). Its credibility depends on showing the trends were parallel *before* the treatment.
- **Interrupted time series (ITS).** A single group, many time points, a sharp treatment at a known date — e.g., an agency's award patterns before vs. after it changed its published review criteria. Assumes the treatment is the only thing that changed at that point.

- **Regression discontinuity (RD).** A treatment assigned by a threshold on a running variable — e.g., a small-business program's eligibility cutoff on firm size. Units just above vs. just below the cutoff are near-randomly assigned. This field's *thresholds* — the 0.42 line, the eligibility cutoffs — are natural RD designs waiting for a researcher.
- **Instrumental variables.** A third variable that affects the treatment but not the outcome except through the treatment. Hard to find, powerful when justified.
- **Natural experiments.** Policy changes, program openings and closings, agency reorganizations — the field is full of them, and each is a design opportunity.

The seminar's discipline for causal claims: **name the identifying assumption.** Every causal claim rests on an assumption the data cannot fully verify; the researcher's job is to say what it is and why it is plausible.

Preregistration and reproducibility

The field's young empirical base will only be trusted if its studies are honest. The course installs the two habits that protect honesty:

- **Preregistration.** Writing the analysis plan — the model, the covariates, the outcome definitions, the sensitivity analyses — *before* seeing the results, and dating it. This prevents the field's most common self-deception: choosing the analysis that flatters the finding.
- **Reproducible methods writing.** A methods section a stranger could execute: data source, sample construction, variable definitions, model specifications, robustness checks. The dissertation's analysis must be reproducible from its documented data and methods even when the data itself is confidential.

Seminar activity. Each student brings their design from DL 901 and names: the causal claim (or its absence), the identifying assumption, the two most serious bias traps for their study, and their mitigations. The room critiques.

Reading. doctrine/04 (gates as the field's natural interventions); doctrine/05.

Deliverable. The **preregistered analysis plan** for the student's study: outcome, predictors, model, covariates, planned robustness checks, and the stated identifying assumption.

Assessment

Layer	Weight	What it is
Movement memos	40%	The four movement deliverables (foundations, model specification, calibration, preregistered plan). Graded on correctness, completeness, doctrine use, and clarity.
Data analysis practicum	40%	On a public, real dataset, the student executes the preregistered plan: fits the model, reports the prediction-quality measures, runs the robustness checks, and writes the methods section. The analysis is checked for reproducibility.
Seminar participation	20%	Honest quantitative critique of classmates' analyses — the peer-review posture the field will demand of its published research.

The course in two sentences

The field's central empirical act is comparing a predicted probability to a real award — and doing that honestly is harder than it looks, because who gets pursued, who resolves, and who wins are all chosen, not random. This course gives you the statistics to make the comparison — and the discipline to know when the evidence will not support the claim you wanted to make.

Reflection questions

1. What is the identifying assumption behind your strongest causal claim, and how would a skeptic attack it?
2. What would a well-calibrated pWin estimate look like in your data — and what would an overconfident one look like?
3. If a stranger ran your analysis from your methods section alone, would they get your numbers?