# A Statistics Primer for Fairness in Machine Learning

Angelina Wang

Last Updated: May 8, 2026

**What this is.** A practical statistics primer and quick reference for fairness evaluation and auditing. It targets common situations that come up in fairness work and common statistical mistakes people make.
**What this is not.** A comprehensive or exhaustive statistics text; an overview of fairness metrics.

## Contents

## 1 Quick Summary

**If you only remember 5 things:**

1. **Always report subgroup size $n$ and uncertainty.** Include a confidence interval (or bootstrap interval) with your point estimate.
2. **Compare groups directly.** If you care about a gap, compute a CI for the *gap*—not two separate CIs and eyeballing overlap.
3. **Be careful with lots of tests.** Many groups $\times$ many metrics $\Rightarrow$ some "significant" results will appear by chance.
4. **Watch for non-independence.** Templates and other repeated structures reduce the effective sample size; use cluster-aware uncertainty when possible.
5. **Don't arbitrarily drop outliers.** Check whether extreme values concentrate in protected groups and report sensitivity analyses.

Table 1: Quick glossary

| Term | Plain-English meaning |
|---|---|
| Estimate | The number you computed from your dataset (e.g., 12/40). |
| Standard error (SE) | How much your estimate would vary if you repeated the same audit on new samples (and a building block for CIs and tests). |
| Confidence interval (CI) | A range meant to reflect sampling uncertainty (e.g., a 90%, 95%, or 99% CI for a mean, a rate, or a group gap). Often built from the SE: CI $\approx$ estimate $\pm$ (critical value)$\times$SE (for 95%, often close to $\pm 2\times$SE), or from bootstrap quantiles. |
| $p$-value | Under a "no difference" assumption, how surprising your data are; *not* the probability the null is true. |
| Multiple hypothesis testing | If you try lots of comparisons, some will look "significant" by luck, so you should correct for this. |

## 2 Keywords

Table 1 defines the key terms; a note that there is related literature on how the estimate may be a poor operationalization of the construct of interest [1, 2]; that is not covered here.

### 2.1 A quick map of common fairness quantities

Many group fairness metrics are functions of subgroup rates:

- **Selection rate**: $\hat{p}_g = \frac{\#\{\hat{Y}=1, G=g\}}{\#\{G=g\}}$.
- **True positive rate (TPR)**: $\widehat{\mathrm{TPR}}_g = \frac{\#\{\hat{Y}=1, Y=1, G=g\}}{\#\{Y=1, G=g\}}$.
- **False positive rate (FPR)**: $\widehat{\mathrm{FPR}}_g = \frac{\#\{\hat{Y}=1, Y=0, G=g\}}{\#\{Y=0, G=g\}}$.
- **Disparity measures**: differences (e.g., $\hat{p}_A - \hat{p}_B$) or ratios (e.g., $\hat{p}_A / \hat{p}_B$).

When denominators are small, all of these can have wide uncertainty. This mini-map is only here so later sections have names to refer to; you do not need to memorize it.

## 3 Means, Medians, and Outliers

Means and medians are two common statistics reported to summarize group quantities.

### 3.1 Means vs. medians: what question are you answering?

The mean is sensitive to extreme values; the median is robust to them. Neither is "better" in general. In fairness contexts, the choice encodes a value judgment about what matters:

- **Means** emphasize aggregate resource allocation and can reflect tail harms (e.g., extremely long delays).
- **Medians/quantiles** reflect the "typical" experience and can hide rare but severe harms.

> **Fairness lens.** "Outliers" are often *people experiencing the worst outcomes.* Automatically dropping them can erase disparate harm. If you filter at all, do it only for clear data-quality failures (e.g., duplicates, impossible values, logging bugs), and consider showing how the conclusions change with and without the rule.

### 3.2 A concrete example: medians can hide tail harms

Suppose two groups have similar median wait times for a service, but Group $A$ has a small fraction of users who experience multi-hour delays due to a routing failure that disproportionately affects them. A median comparison may show "no disparity," while the mean or a high-quantile metric (e.g., the 90th percentile) reveals a meaningful harm concentration. In fairness work, it is often valuable to report both a typical-experience metric (median) and a tail metric (e.g., 90th/95th percentile), with uncertainty for each.

### 3.3 Robust ways to summarize distributions

- **Box-and-whisker plots (by group)** to visualize medians, spread, and extreme values.
- **Histograms or ECDF plots (by group)** when you want to see the whole distribution shape.
- **Quantile gaps** (e.g., difference in 90th/95th percentile latency) to focus on tail harms.
- **Sensitivity analysis** (show how a disparity changes as you vary a reasonable robustness choice).

## 4 Confidence Intervals

> **Practical warning.** A confidence interval only quantifies uncertainty for a specific target quantity under a specific sampling story. If your data are a filtered log, a hand-labeled audit, or a biased sample, the interval does *not* automatically fix that.

### 4.1 What a confidence interval is (and is not)

A $(1-\alpha)$ confidence interval for a parameter $\theta$ is a procedure that, under repeated sampling, contains the true $\theta$ with probability $(1-\alpha)$. A practical way to read a "95% CI" is:

> If we repeated the same data-collection process many times and rebuilt the interval each time, about 95% of those intervals would contain the true value.

**Common misreadings:** a 95% CI is not the probability the parameter lies in this particular interval, and it is not a range that contains 95% of individual data points.

### 4.2 CI for a mean (continuous outcomes)

For i.i.d. data $X_1, \ldots, X_n$ with sample mean $\bar{X}$ and sample standard deviation $s$, a common $(1-\alpha)$ CI for the population mean $\mu$ is

$$\bar{X} \pm t_{n-1,1-\alpha/2} \frac{s}{\sqrt{n}},$$

where $t_{n-1,1-\alpha/2}$ is a Student-$t$ quantile.

**Assumptions and small $n$.** The $t$-interval is exact if the data are normally distributed; otherwise it is an approximation that typically improves with larger $n$ (via the central limit theorem). With small subgroup sizes, or with heavy tails/skew, consider robust summaries (Section 3) or bootstrap intervals (below).

### 4.3 CI for a proportion (rates and error rates)

If $X \sim \text{Binomial}(n, p)$ and $\hat{p} = X/n$, a common but fragile interval is the "Wald" interval $\hat{p} \pm z_{1-\alpha/2}\sqrt{\hat{p}(1-\hat{p})/n}$. It can behave badly at small $n$ or when $\hat{p}$ is near 0 or 1.

**Recommendation.** Use a more stable binomial CI for rates, such as a Wilson interval (or an exact/Clopper–Pearson interval when you want conservatism). The precise choice is less important than (i) avoiding the Wald interval in small-$n$ and (ii) reporting the method used.

### 4.4 Bootstrap confidence intervals (general-purpose)

For both quantities that fit into the categories above, as well as quantities that are more complex without clear analytical ways to compute confidence intervals, a bootstrap can be a good default:

1. For each bootstrap replicate $b = 1, \ldots, B$, resample the data *with replacement*, then recompute the metric $\hat{\theta}^{(b)}$.
2. Use the empirical quantiles of $\{\hat{\theta}^{(b)}\}$ to form an interval (e.g., the 2.5th and 97.5th percentiles for 95%).

**Resampling scheme matters.** If your data have dependence (e.g., multiple records per person, time series, clustered sampling), naive bootstrap resampling of individual rows can understate uncertainty. You may need a cluster bootstrap (resample people, then their rows) or a block bootstrap (resample time blocks).

### 4.5 Quick reference: choosing an interval method

Refer to Table 2, remembering that bootstrapping is an option as well.

Table 2: Choosing an interval method

| Quantity | Reasonable default CI | Notes for fairness work |
|---|---|---|
| Mean (continuous) | $t$-interval | Heavy tails are common; consider medians/quantiles too. |
| Proportion/rate | Wald; Wilson | Always report denominators; Wilson handles small samples better than Wald. |
| Difference of means | Welch | Prefer Welch by default (it does not assume equal variances). |

### 4.6 When examples are not independent: templates and "clustered" data

Many fairness benchmarks and audits contain **clusters** of very similar examples. A common pattern is **templates**: you write a small number of prompt templates and then instantiate each template for many demographic groups (or many names, locations, etc.). This is useful for controlled comparisons, but it creates dependence.

> **Why this matters.** If you have 10 templates and each one is repeated 100 times, you do *not* have 1000 fully independent data points. Treating them as independent can make confidence intervals too narrow and can make disparities look more "certain" than they really are.

**Practical approaches:**

- **Aggregate by template:** compute your metric per template, then summarize across templates (and report how many templates you had).
- **Report cluster counts:** report how many templates (or people) you have, not just how many total rows.
- **Cluster bootstrap:** resample templates (clusters) with replacement, not individual rows.

## 5 Comparing Groups: Differences, Ratios, and the "CI Overlap" Trap

### 5.1 Report the comparison you care about

In fairness contexts, the primary quantity is often a *difference* or *ratio* between groups:

$$\Delta = \theta_A - \theta_B \qquad \text{or} \qquad R = \theta_A / \theta_B.$$

If you care about $\Delta$, compute an estimate $\hat{\Delta}$ and a confidence interval for $\Delta$ directly.

### 5.2 The CI-overlap trap: don't eyeball two separate intervals

In public discussions of disparity, people often compare two subgroup confidence intervals and conclude there is (or is not) a meaningful difference based on whether the intervals overlap. This is unreliable.

**Rule.** If your question is "how different are these two groups?", compute an interval for the *difference* (or ratio). Do not use overlap of two separate intervals as your decision rule.

This theme appears often in fairness measurement discussions; see Lum et al. for a related perspective [3].

**Why overlap is a bad test**

Let $\hat{\theta}_A$ and $\hat{\theta}_B$ be (approximately) independent, with standard errors $\mathrm{SE}_A$ and $\mathrm{SE}_B$. A significance-style comparison uses the standard error of the difference:

$$\mathrm{SE}(\hat{\theta}_A - \hat{\theta}_B) = \sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2}.$$

But "do the 95% CIs overlap?" informally compares $1.96(\mathrm{SE}_A + \mathrm{SE}_B)$, which is larger than $1.96\sqrt{\mathrm{SE}_A^2 + \mathrm{SE}_B^2}$. Therefore, two 95% intervals can overlap even when the difference is statistically distinguishable at the 5% level.

### 5.3 Difference of means: a simple approximate interval

If $A$ and $B$ are independent samples with means $\bar{X}_A, \bar{X}_B$, variances $s_A^2, s_B^2$, sizes $n_A, n_B$, then a simple approximate 95% CI for the mean difference is

$$(\bar{X}_A - \bar{X}_B) \pm t^* \sqrt{\frac{s_A^2}{n_A} + \frac{s_B^2}{n_B}}.$$

Here $t^*$ is the critical value for your chosen confidence level; for a 95% CI it is often close to 2, especially when sample sizes are not small.

### 5.4 Difference of rates (risk difference)

For rates $\hat{p}_A = X_A/n_A$ and $\hat{p}_B = X_B/n_B$, an approximate CI for $\Delta = p_A - p_B$ is

$$(\hat{p}_A - \hat{p}_B) \pm z_{1-\alpha/2} \sqrt{\frac{\hat{p}_A(1-\hat{p}_A)}{n_A} + \frac{\hat{p}_B(1-\hat{p}_B)}{n_B}}.$$

For small denominators, this normal approximation can be shaky, so a bootstrap interval for the gap may be easier to justify.

### 5.5 Worked example: a selection-rate gap with small subgroup $n$

Suppose an audit of a decision system yields:

- Group $A$: $n_A = 40$ cases, $X_A = 12$ selected, so $\hat{p}_A = 0.30$.
- Group $B$: $n_B = 200$ cases, $X_B = 80$ selected, so $\hat{p}_B = 0.40$.

The estimated disparity (risk difference) is $\hat{\Delta} = \hat{p}_A - \hat{p}_B = -0.10$ (a 10 percentage-point gap).

Using the normal approximation for illustration,

$$\mathrm{SE}(\hat{\Delta}) \approx \sqrt{\frac{0.30 \cdot 0.70}{40} + \frac{0.40 \cdot 0.60}{200}} \approx 0.080,$$

so an approximate 95% CI is

$$-0.10 \pm 1.96(0.080) \approx [-0.26,\ 0.06].$$

This interval is wide: with this data alone you cannot rule out "no gap," but you also cannot rule out a substantial gap. In a fairness setting, this is a cue to (i) collect more evidence, (ii) avoid overconfident claims either way, and (iii) consider whether a precautionary response is warranted given potential harm.

## 6 Hypothesis Tests: $t$-Tests and Permutation Tests

The previous section is about *estimation*: report the group gap you care about (difference or ratio) with a confidence interval. This section is about *testing*: producing a $p$-value for a narrow question like "if the true gap were exactly 0, how surprising is my data?"

> **Note:** In fairness work, effect sizes and uncertainty intervals are often more aligned with decisions than a binary "significant / not significant" conclusion.

### 6.1 Welch two-sample $t$-test (difference in means)

Use a $t$-test when your primary question is a **difference in means** between two groups (or between two conditions).

- **Typical fairness use:** "Is the *average* error severity (or average wait time, or average score) different for Group A vs. Group B?"
- **Null hypothesis:** $\mu_A = \mu_B$.
- **When it makes sense:**
  - the outcome is roughly continuous, and the *mean* is a meaningful summary,
  - observations are independent within and across groups,
  - subgroup sizes are not extremely tiny (very small $n$ makes both tests and CIs fragile).
- **Default choice:** use **Welch's** $t$-test (it does not assume equal variances).
- **Report alongside the test:** the estimated mean gap and its CI (a $p$-value alone hides magnitude).

### 6.2 Permutation tests (fewer distributional assumptions)

Permutation tests are a flexible option when you want a $p$-value but do not want to lean on normal-approximation assumptions.

- **Core idea:** under a "no group difference" null, group labels are exchangeable; the observed gap should look typical among gaps computed after randomly shuffling labels.
- **Pick a test statistic:** any disparity you care about, e.g., difference in means, difference in medians, or a rate gap.
- **Procedure (two-sided, Monte Carlo):**
  1. Compute the observed statistic $T_{\text{obs}}$ on the real labels.

2. For $b = 1, \ldots, B$, randomly permute the group labels (or permute within strata; see below) and recompute $T^{(b)}$.
3. Estimate $p \approx \frac{1}{B} \sum_{b=1}^{B} \mathbf{1}\{|T^{(b)}| \geq |T_{\text{obs}}|\}$.

- **Fairness caveat (confounding):** if group membership is associated with other factors that drive the outcome, naive label-shuffling creates an unrealistic null. A more defensible approach is often **stratified** permutation (shuffle labels only within matched pairs, templates, time windows, or other strata meant to hold key covariates fixed).
- **Dependence caveat:** if data are clustered (e.g., templates or multiple rows per person), permute at the cluster level (shuffle templates/people, not individual rows) or use a cluster-aware resampling scheme.
- **Report alongside the test:** the same gap with a CI (often bootstrap), so readers can see practical significance.

**Other tests you may encounter (as needed):** Fisher's exact test (tiny $2 \times 2$ counts), two-proportion/$\chi^2$ tests (rate differences), Mann–Whitney/Wilcoxon rank-sum (rank-based distribution shifts), and McNemar's test (paired binary outcomes, e.g., two models on the same cases).

## 7 Multiple Hypothesis Testing in Fairness Audits

### 7.1 Why multiple comparisons are unavoidable

Fairness evaluation typically involves:

- many subgroups (and intersectional subgroups),
- many metrics (selection rate, TPR, FPR, calibration, error severity),
- many thresholds (decision cutoffs),
- many models/hyperparameters (if you iterate).

If you run (say) 100 independent 5%-level tests under a global "no disparity" world, you expect about 5 false positives by chance. **Tip:** In some fields, preregistration is used to reduce "researcher degrees of freedom." You can do the same: preregister (or at least pre-specify) your primary groups/metrics and stopping rules *before* you look at the results.

### 7.2 Two common goals: FWER vs FDR

- **Family-wise error rate (FWER):** controls the probability of *any* false positive in a family of tests. Methods: Bonferroni, Holm. Conservative but simple.
- **False discovery rate (FDR):** controls the expected proportion of false positives among discoveries. A common method is Benjamini–Hochberg (BH). Often more powerful when testing many hypotheses.

**Pragmatic fairness workflow.** Pre-specify a small number of primary comparisons (the disparities you will act on), use a correction or an error-control method for those, and treat the rest as exploratory—but then demand replication or additional evidence before making high-stakes claims.

### 7.3 Benjamini–Hochberg (BH) procedure

Given $m$ $p$-values, sort them $p_{(1)} \le \cdots \le p_{(m)}$. Find the largest $k$ such that $p_{(k)} \le \frac{k}{m} q$ where $q$ is the target FDR (e.g., $q = 0.10$). Declare tests $1, \ldots, k$ "discoveries." (Assumptions matter; BH is most reliable under independence or certain positive dependence conditions.)

## 8 Regression-Based Comparisons and Controls

Regression is useful when you want to compare groups while also accounting for other variables such as case complexity, geography, prompt template, or time period. In fairness work, that changes the question from an overall disparity to a *conditional* disparity: how different are groups after holding the included covariates fixed?

### 8.1 Reading a group coefficient

For a continuous outcome, a simple linear regression is

$$Y_i = \beta_0 + \beta_1 \mathbf{1}\{G_i = A\} + \beta_2 X_i + \varepsilon_i,$$

where $X_i$ is another variable you want to hold fixed.

- $\beta_1$ is the adjusted difference in the mean outcome between Group $A$ and the reference group (e.g., Group B), conditional on $X_i$.
- With more than two groups, include multiple group indicators and interpret each coefficient relative to the chosen reference group. Johfre and Freese point out that the reference category changes which contrast is foregrounded, so state it explicitly and avoid treating a dominant group (e.g., men, White people) as the automatic default [4].

### 8.2 When to control for other variables (the $X_i$ above)

Common reasons to add covariates include making comparisons within the same geography, case type, time period, or prompt template, or adjusting for observable differences in task difficulty. This can be valuable, but it is not automatically the "right" fairness analysis.

> **Fairness caution about controls.** If you control for a variable that lies on the pathway from group membership to the outcome, or for a variable that is itself measured with bias, you can make a real disparity look smaller. Regression does not tell you automatically which controls are appropriate.

**Practical rule:** when possible, report both the raw disparity and the adjusted disparity, then explain what each one means.

### 8.3 Choosing a model for the outcome

- **Linear regression** is a common default for continuous outcomes such as scores, delays, or error severity.
- **Logistic regression** is common for binary outcomes such as selection, approval, or whether an answer is harmful. Its coefficients live on the log-odds scale, so for communication it is often better to report adjusted predicted probabilities or marginal effects in addition to odds ratios.

- **Other generalized linear models** can be appropriate for counts or rates, but the same principle applies: make clear what scale the coefficient lives on and translate it back into a quantity readers can interpret.

### 8.4 Worked example: an adjusted group gap

Suppose you are auditing response quality scores on a 0–10 scale, where larger values mean harsher or less helpful model behavior. Group $A$ prompts tend to be more difficult than the reference group's prompts, so you fit

$$Y_i = \beta_0 + \beta_1 \mathbf{1}\{G_i = A\} + \beta_2 D_i + \varepsilon_i,$$

where $D_i$ is a task-difficulty score.

Imagine the fitted model gives $\hat{\beta}_1 = 0.6$ with a 95% CI of $[0.1,\ 1.1]$, and $\hat{\beta}_2 = 1.4$. A practical reading is:

- holding difficulty fixed, Group $A$ still receives responses that are about 0.6 points harsher on average than the reference group;
- the positive coefficient on difficulty means harder prompts tend to increase harshness scores for everyone;
- if the raw mean gap had been 1.5 points, the regression would suggest that some, but not all, of that raw disparity is associated with prompt difficulty.

This is the main appeal of regression in fairness work: it lets you ask whether a group gap remains after accounting for a variable you think should be held fixed.

### 8.5 Interactions: does the gap change across settings?

If you want to know whether the group gap changes with another variable, add an interaction:

$$Y_i = \beta_0 + \beta_1 \mathbf{1}\{G_i = A\} + \beta_2 X_i + \beta_3 \mathbf{1}\{G_i = A\} X_i + \varepsilon_i.$$

Here $\beta_3$ captures whether the association between $X_i$ and the outcome differs for Group $A$ relative to the reference group. This is useful when disparities may vary by task difficulty, age, time, or another feature.

### 8.6 Clustered data, fixed effects, and mixed-effects models

Many fairness datasets contain repeated structure: multiple decisions by the same reviewer, multiple prompts from the same template, or repeated observations for the same person or site.

- **Fixed effects** add a separate intercept for each template, person, site, or time block when you want comparisons *within* those units. In effect, you ask whether a group gap remains among cases that share the same template, reviewer, or site.
- **Mixed-effects models** also account for repeated structure, but instead of adding a separate free coefficient for every unit, they model unit-to-unit variation as coming from a shared distribution. This is often useful when you have many templates or sites and want partial pooling rather than a long list of fixed-effect indicators. A simple random-intercept model is

  $$Y_{ij} = \beta_0 + \beta_1 \mathbf{1}\{G_{ij} = A\} + \beta_2 X_{ij} + u_j + \varepsilon_{ij},$$

  where $u_j$ is a template-, person-, or site-specific intercept.

**When repeated structure matters, do not rely on row-level independence.** If examples share templates, annotators, reviewers, or users, use a model that represents that structure explicitly, such as fixed effects or mixed effects, and report how many clusters you actually had.

Mixed-effects models are often attractive when you have many clusters and want partial pooling. Fixed effects are often easier to explain when the main goal is to compare groups within a known set of templates, reviewers, or sites.

### 8.7 Common mistakes to avoid

- **Treating an adjusted coefficient as causal.** Regression adjustment does not, by itself, show what would happen under an intervention.
- **Forgetting the reference group.** A coefficient on group membership is always a contrast against some baseline, so name that baseline explicitly.
- **Controlling for the wrong variable.** If a covariate is downstream of group membership, or is itself measured with bias, adjustment can hide a meaningful disparity.
- **Reporting only raw coefficients from nonlinear models.** For logistic regression in particular, readers usually understand adjusted probabilities or marginal effects better than log-odds.
- **Using a very elaborate model with tiny subgroup counts.** Adjustment can stabilize some comparisons, but sparse data can still make subgroup coefficients noisy, fragile, and more prone to overfitting.
- **Over-reading $R^2$.** In a linear regression, $R^2$ is the fraction of variation in the outcome explained by the predictors in-sample. A high $R^2$ does not mean the group coefficient is fair or causal, and a low $R^2$ does not mean the group comparison is unimportant; fairness questions are often about a specific disparity, not overall fit.

## Acknowledgments

Thank you to Vyoma Raman for feedback. I used OpenAI Codex to help draft this.

## References

[1] Abigail Z. Jacobs and Hanna Wallach. Measurement and fairness. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, 2021.

[2] Ian Lundberg and Rebecca Johnson and Brandom M. Stewart. What is your estimand? Defining the target quantity connects statistical evidence to theory. In *American Sociological Review*, 2021.

[3] Kristian Lum, Yunfeng Zhang, and Amanda Bower. De-biasing "bias" measurement. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency (FAccT '22)*, 2022.

[4] Sasha Shen Johfre and Jeremy Freese. Reconsidering the reference category. In *Sociological Methodology*, 51(2):253–269, 2021.