

## 1 PROBABILITY DISTRIBUTIONS

*PDF.* (Probability Density Function) describes the likelihood of a continuous random variable taking a specific value over an interval.

$$f(x) \geq 0 \quad \forall x, \quad \int_{-\infty}^{\infty} f(x)dx = 1.$$

The PDF represents the area under the curve, and if  $b = a$ , then  $P(a \leq X \leq b) = 0$ , which is known as the zero-point problem.

*CDF.* (Cumulative Distribution Function) represents the probability that a variable is less than or equal to a given value:

$$F(x) = P(X \leq x) = \int_{-\infty}^x f(t)dt, f(x) = \frac{d}{dx} F(x)$$

The CDF is non-decreasing, and it satisfies the limits:

$$\lim_{x \rightarrow -\infty} F(x) = 0, \lim_{x \rightarrow \infty} F(x) = 1.$$

*Expectation and Variance.*

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx, E[X^2] = \int_{-\infty}^{\infty} x^2f(x)dx,$$

$$\text{Var}(X) = E[X^2] - (E[X])^2, \sigma_X = \sqrt{\text{Var}(X)}$$

## 2 HYPOTHESIS TESTING

*One-Sample Mean Test.*

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}}, t = \frac{\bar{X} - \mu}{s/\sqrt{n}}, s = \sqrt{\frac{\sum(X_i - \bar{X})^2}{n-1}}$$

*Two-Sample Mean Test.* 2 samples are random and come from 2 distinct populations that are independent. Populations are also normally distributed. The first row is if population variance is known (Z-test) and if equal (pooled) variances (T-test).

$$Z = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}},$$

$$t = \frac{(\bar{X}_1 - \bar{X}_2) - \Delta}{S_p \sqrt{\frac{1}{n_1} + \frac{1}{n_2}}}, S_p = \sqrt{\frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2}}$$

$$\text{dof} = n_1 + n_2 - 2$$

*Probability of a Type I error (False Positive):*

$$\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true})$$

A lower value of  $\alpha$  indicates a more stringent criterion for rejecting the null hypothesis, the significance level.

*Probability of a Type II error (False Negative):*

$$\beta = P(\text{Fail to reject } H_0 \mid H_0 \text{ is false})$$

Indicates how often a test fails to detect a true effect.

*Power of the test (Sensitivity):*

$$\pi = 1 - \beta = P(\text{Reject } H_0 \mid H_0 \text{ is false})$$

This measures the ability of the test to detect an effect when one truly exists. A higher power ( $\pi$ ) means the test is more likely to identify a true effect, to reject a false null hypothesis correctly.

*Sample Size (n):*

$$n = \left( \frac{(Z_\alpha + Z_\beta) \cdot \sigma}{\delta} \right)^2, \left( \frac{(Z_{\alpha/2} + Z_\beta) \cdot \sigma}{\delta, \text{effect size}} \right)^2$$

Assume to keep the same power, but reduce sample size  $n$ :

- Increase  $\delta$ , if the test is more effective to produce a clearer difference, need fewer samples to identify the difference
- Use 1-sided if possible to care only about 1 direction, this requires fewer samples

- Increase  $\alpha$  which in turn decreases  $Z_\alpha$ , which requires fewer samples, makes it easier to detect statistically significant differences to reject the null hypothesis
- Reduce  $\sigma$  by using more precise measurements, decreases SE, allowing us to identify the effect with smaller samples
- Note: Increasing  $\alpha$  reduces  $Z_\alpha$ .
- Note: Increasing  $\beta$  (decreasing power) reduces  $Z_\beta$ .

## 3 REGRESSION ANALYSIS

*Sum of Squares Decomposition.*

$$SST/TSS \text{ (Total)} = \sum (Y_i - \bar{Y})^2$$

(Total variance in dependent var Y)

$$SSR/ESS \text{ (SS Regression/ Explained SS)} = \sum (\hat{Y}_i - \bar{Y})^2$$

(Explained variance in Y by regression model)

$$SSE/RSS \text{ (SS Error/Residual SS)} = \sum (Y_i - \hat{Y}_i)^2$$

$$\text{(Unexplained variance in Y)} = \sum (e_i)^2$$

*R-Squared and Adjusted R-Squared.* Coefficient of determination (model fit): Proportion of variation in Y explained by the model. If 0 implies the model does not explain anything. If 1 implies a perfect fit. An increasing number of predictors will always increase or keep R-squared the same even if a new predictor is not useful. Adjusted R-Squared accounts for the number of predictors to prevent overfitting. Only increases if new predictor improves model and is not by chance, else adjusted R-squared drops.

$$R^2 = 1 - \frac{RSS}{TSS} = \frac{ESS}{TSS}, \text{Adjusted } R^2 = 1 - \left( \frac{(1 - R^2)(n - 1)}{n - k - 1} \right)$$

## 4 HETEROSKEDASTICITY

Heteroskedasticity occurs when the variance of residuals in a regression model is not constant, violating the assumption of homoskedasticity in Ordinary Least Squares (OLS).

$$\text{Var}(\epsilon_i | X_i) \neq \sigma^2 = f(x_i)$$

*Effects of Heteroskedasticity.*

- (1) Inefficient OLS estimates (not Best Linear Unbiased Estimators)
- (2) Incorrect SEs, leading to unreliable hypothesis tests
- (3) Biased confidence intervals and p-values

*Detection Methods.*

- (1) **Residual Plots:** Check if residuals show a funnel shape.
- (2) **White Test:** Regressing squared residuals on independent variables and their squares, then checking significance.

*Solutions.*

- (1) Log or Box-Cox transformations can stabilize the variance.
- (2) Use White's robust standard errors, by replacing the usual OLS variance-covariance matrix  $(X'X)^{-1}\sigma^2$  with  $(X'X)^{-1}(X'\Omega X)(X'X)^{-1}$ , where  $\Omega$  is a diagonal matrix of squared residuals. Best practise is to use robust SE regardless.
- (3) Weights observations to adjust for variance differences.

## 5 REGRESSION COEFFICIENTS

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \dots + \beta_k X_k + \epsilon,$$

- $\beta_0$  (Intercept): Represents the expected value of Y when all independent variables ( $X_i$ ) are equal to zero.
- $\beta_i$  (Slope Coefficient): Represents the expected change in Y for a one-unit increase in  $X_i$ , holding all other X variables constant.

**Dummy Variables.** Represent categorical data in regression. For a variable with  $k$  categories, we create  $k - 1$  dummy variables to avoid perfect collinearity. The omitted category serves as the **reference category**, against which the effects of other categories are compared.

**Interaction Terms.** Interaction terms capture situations where one independent variable's effect depends on another's value. An interaction term between two variables  $X_1$  and  $X_2$  is represented as:

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \beta_3 (X_1 \cdot X_2) + \epsilon.$$

Here,  $\beta_3$  represents the interaction effect. If significant, it indicates that the relationship between  $X_1$  and  $Y$  depends on the value of  $X_2$ , and vice versa.

## 6 OMITTED VARIABLE BIAS (OVb)

OVb occurs when a relevant variable is excluded from a regression model, leading to biased coefficient estimates.

$$Y = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \epsilon$$

If  $X_2$  is omitted, the estimated model is:

$$Y = \alpha_0 + \alpha_1 X_1 + u$$

The bias in  $\alpha_1$  is:

$$\text{Bias}(\alpha_1) = \beta_2 \cdot \rho_{X_1, X_2}$$

where  $\rho_{X_1, X_2}$  is the correlation between  $X_1$  and  $X_2$ .

**Implications of OVb.**

- (1) If  $\beta_2$  and  $\rho_{X_1, X_2}$  same sign,  $X_1$ 's coefficient is **overestimated**.
- (2) If they have opposite signs, it is **underestimated**.

## 7 LOCAL AVERAGE TREATMENT EFFECT

LATE is the causal effect of a treatment on *compliers* in an Instrumental Variables (IV) framework. It is estimated when treatment assignment is not strictly followed.

- (1) **Compliers:** Take treatment if assigned.
- (2) **Always Takers:** Always take treatment.
- (3) **Never Takers:** Never take treatment.
- (4) **Defiers:** Do the opposite of the assignment (assumed rare).

**Wald Estimator for LATE.** When an instrument  $Z$  (randomly assigned) influences whether an individual actually receives a treatment  $D$ , but some individuals do not comply with their assignment, we often estimate the Local Average Treatment Effect (LATE) with the Wald Estimator. Let  $Y$  be the outcome of interest.

$$\begin{aligned} \text{LATE} &= \frac{E[Y | Z = 1] - E[Y | Z = 0]}{E[D | Z = 1] - E[D | Z = 0]} \\ &= \frac{\text{ITT}_Y \text{ (Intent-to-Treat effect on the outcome)}}{\text{ITT}_D \text{ (Intent-to-Treat effect on actual treatment)}} \end{aligned}$$

Suppose half of the eligible households are randomly offered a Green Voucher ( $Z = 1$ ) to subsidize the purchase of solar panels, and the other half are not offered the voucher ( $Z = 0$ ). Let  $D$  indicate whether the household installs solar panels, and let  $Y$  be the (annual) reduction in electricity expenses.

$$\begin{aligned} E[Y | Z = 1] &= 2800 \text{ (Avg reduction for those offered the voucher)} \\ E[Y | Z = 0] &= 2100 \text{ (Avg reduction for those not offered the voucher)} \\ E[D | Z = 1] &= 0.50 \text{ (Those offered voucher, \% actually install panels)} \\ E[D | Z = 0] &= 0.10 \text{ (Those not offered voucher, \% install panels still)} \end{aligned}$$

$$\text{LATE} = 1750$$

Among households whose decision to install solar panels *depends* on whether they receive the voucher (the compliers), effect of installing solar panels is an additional \$1750 of electricity expense reduction per year.

## 8 DIFFERENCE-IN-DIFFERENCES (DID)

DiD estimates causal results by comparing differences in outcomes before and after a policy change or treatment between the treatment group and the control group:

$$DID = (Y_{T,Post} - Y_{T,Pre}) - (Y_{C,Post} - Y_{C,Pre})$$

where  $T$  and  $C$  are treatment and control groups respectively.

- Assumes both groups follow parallel trends in the absence of treatment.
  - Compare historical trends of both groups before treatment.
- Assumes no other factors (besides treatment) caused the observed change.
  - Use additional control variables.
- Groups have similar characteristics.

**Example Calculation.** Before treatment and post-treatment:

- Treatment group: 80%  $\rightarrow$  75%
- Control group: 85%  $\rightarrow$  82%

$$DID = (75 - 80) - (82 - 85) = -2$$

**Interpretation:** The estimated effect of treatment is a 2% drop.

**Regression Specification.**

$$\begin{aligned} Y &= \beta_0 + \beta_1 (\text{Post}) + \beta_2 (\text{Treatment}) \\ &\quad + \beta_3 (\text{Post} \times \text{Treatment}) + \epsilon + \gamma X_i \end{aligned}$$

where:

- **Post:** 0 = Pre-treatment, 1 = Post-treatment
- **Treatment:** 0 = Control group, 1 = Treatment group
- $\beta_3$  captures the **DiD effect**.
- Additional covariates  $X_i$  can be included to account for confounders, reduce variability, and increase the precision of  $\beta_3$ .

$$\begin{aligned} \text{Revenue} &= 500 + 20(\text{Post}) - 50(\text{Treatment}) \\ &\quad - 30(\text{Treatment} \times \text{Post}) + \epsilon \end{aligned}$$

**Interpretation:**

- 500: Base revenue for the control group before treatment.
- 20: Increase in revenue after treatment common to both groups.
- -50: Baseline difference between both groups before treatment.
- -30: The DiD effect—treatment led to a 30 reduction in revenue for treated subjects relative to control subjects post-treatment.

## 9 DATA SNOOPING

Data snooping occurs when a dataset is analyzed repeatedly until statistically significant patterns emerge, leading to overfitting and false positives.

**Common Issues.**

- (1) Multiple hypothesis testing without corrections.
- (2) Repeated tuning of a model without proper validation.
- (3) Look-ahead bias: Using future data for predictions.

**Prevention Strategies.**

- (1) Use a proper **train-test split** to prevent leakage.
- (2) Apply **cross-validation** (e.g., K-Fold) to ensure generalization.
- (3) Adjust for multiple testing using **Bonferroni correction** or **False Discovery Rate (FDR)**.
- (4) Pre-register hypotheses before testing.

**STANDARD NORMAL DISTRIBUTION: Table Values Represent AREA to the LEFT of the Z score.**

<b>Z</b>	<b>.00</b>	<b>.01</b>	<b>.02</b>	<b>.03</b>	<b>.04</b>	<b>.05</b>	<b>.06</b>	<b>.07</b>	<b>.08</b>	<b>.09</b>
<b>-3.9</b>	.00005	.00005	.00004	.00004	.00004	.00004	.00004	.00004	.00003	.00003
<b>-3.8</b>	.00007	.00007	.00007	.00006	.00006	.00006	.00006	.00005	.00005	.00005
<b>-3.7</b>	.00011	.00010	.00010	.00010	.00009	.00009	.00008	.00008	.00008	.00008
<b>-3.6</b>	.00016	.00015	.00015	.00014	.00014	.00013	.00013	.00012	.00012	.00011
<b>-3.5</b>	.00023	.00022	.00022	.00021	.00020	.00019	.00019	.00018	.00017	.00017
<b>-3.4</b>	.00034	.00032	.00031	.00030	.00029	.00028	.00027	.00026	.00025	.00024
<b>-3.3</b>	.00048	.00047	.00045	.00043	.00042	.00040	.00039	.00038	.00036	.00035
<b>-3.2</b>	.00069	.00066	.00064	.00062	.00060	.00058	.00056	.00054	.00052	.00050
<b>-3.1</b>	.00097	.00094	.00090	.00087	.00084	.00082	.00079	.00076	.00074	.00071
<b>-3.0</b>	.00135	.00131	.00126	.00122	.00118	.00114	.00111	.00107	.00104	.00100
<b>-2.9</b>	.00187	.00181	.00175	.00169	.00164	.00159	.00154	.00149	.00144	.00139
<b>-2.8</b>	.00256	.00248	.00240	.00233	.00226	.00219	.00212	.00205	.00199	.00193
<b>-2.7</b>	.00347	.00336	.00326	.00317	.00307	.00298	.00289	.00280	.00272	.00264
<b>-2.6</b>	.00466	.00453	.00440	.00427	.00415	.00402	.00391	.00379	.00368	.00357
<b>-2.5</b>	.00621	.00604	.00587	.00570	.00554	.00539	.00523	.00508	.00494	.00480
<b>-2.4</b>	.00820	.00798	.00776	.00755	.00734	.00714	.00695	.00676	.00657	.00639
<b>-2.3</b>	.01072	.01044	.01017	.00990	.00964	.00939	.00914	.00889	.00866	.00842
<b>-2.2</b>	.01390	.01355	.01321	.01287	.01255	.01222	.01191	.01160	.01130	.01101
<b>-2.1</b>	.01786	.01743	.01700	.01659	.01618	.01578	.01539	.01500	.01463	.01426
<b>-2.0</b>	.02275	.02222	.02169	.02118	.02068	.02018	.01970	.01923	.01876	.01831
<b>-1.9</b>	.02872	.02807	.02743	.02680	.02619	.02559	.02500	.02442	.02385	.02330
<b>-1.8</b>	.03593	.03515	.03438	.03362	.03288	.03216	.03144	.03074	.03005	.02938
<b>-1.7</b>	.04457	.04363	.04272	.04182	.04093	.04006	.03920	.03836	.03754	.03673
<b>-1.6</b>	.05480	.05370	.05262	.05155	.05050	.04947	.04846	.04746	.04648	.04551
<b>-1.5</b>	.06681	.06552	.06426	.06301	.06178	.06057	.05938	.05821	.05705	.05592
<b>-1.4</b>	.08076	.07927	.07780	.07636	.07493	.07353	.07215	.07078	.06944	.06811
<b>-1.3</b>	.09680	.09510	.09342	.09176	.09012	.08851	.08691	.08534	.08379	.08226
<b>-1.2</b>	.11507	.11314	.11123	.10935	.10749	.10565	.10383	.10204	.10027	.09853
<b>-1.1</b>	.13567	.13350	.13136	.12924	.12714	.12507	.12302	.12100	.11900	.11702
<b>-1.0</b>	.15866	.15625	.15386	.15151	.14917	.14686	.14457	.14231	.14007	.13786
<b>-0.9</b>	.18406	.18141	.17879	.17619	.17361	.17106	.16853	.16602	.16354	.16109
<b>-0.8</b>	.21186	.20897	.20611	.20327	.20045	.19766	.19489	.19215	.18943	.18673
<b>-0.7</b>	.24196	.23885	.23576	.23270	.22965	.22663	.22363	.22065	.21770	.21476
<b>-0.6</b>	.27425	.27093	.26763	.26435	.26109	.25785	.25463	.25143	.24825	.24510
<b>-0.5</b>	.30854	.30503	.30153	.29806	.29460	.29116	.28774	.28434	.28096	.27760
<b>-0.4</b>	.34458	.34090	.33724	.33360	.32997	.32636	.32276	.31918	.31561	.31207
<b>-0.3</b>	.38209	.37828	.37448	.37070	.36693	.36317	.35942	.35569	.35197	.34827
<b>-0.2</b>	.42074	.41683	.41294	.40905	.40517	.40129	.39743	.39358	.38974	.38591
<b>-0.1</b>	.46017	.45620	.45224	.44828	.44433	.44038	.43644	.43251	.42858	.42465
<b>-0.0</b>	.50000	.49601	.49202	.48803	.48405	.48006	.47608	.47210	.46812	.46414



# t Table

cum. prob	t <sub>.50</sub>	t <sub>.75</sub>	t <sub>.80</sub>	t <sub>.85</sub>	t <sub>.90</sub>	t <sub>.95</sub>	t <sub>.975</sub>	t <sub>.99</sub>	t <sub>.995</sub>	t <sub>.999</sub>	t <sub>.9995</sub>
one-tail	0.50	0.25	0.20	0.15	0.10	0.05	0.025	0.01	0.005	0.001	0.0005
two-tails	1.00	0.50	0.40	0.30	0.20	0.10	0.05	0.02	0.01	0.002	0.001
df											
1	0.000	1.000	1.376	1.963	3.078	6.314	12.71	31.82	63.66	318.31	636.62
2	0.000	0.816	1.061	1.386	1.886	2.920	4.303	6.965	9.925	22.327	31.599
3	0.000	0.765	0.978	1.250	1.638	2.353	3.182	4.541	5.841	10.215	12.924
4	0.000	0.741	0.941	1.190	1.533	2.132	2.776	3.747	4.604	7.173	8.610
5	0.000	0.727	0.920	1.156	1.476	2.015	2.571	3.365	4.032	5.893	6.869
6	0.000	0.718	0.906	1.134	1.440	1.943	2.447	3.143	3.707	5.208	5.959
7	0.000	0.711	0.896	1.119	1.415	1.895	2.365	2.998	3.499	4.785	5.408
8	0.000	0.706	0.889	1.108	1.397	1.860	2.306	2.896	3.355	4.501	5.041
9	0.000	0.703	0.883	1.100	1.383	1.833	2.262	2.821	3.250	4.297	4.781
10	0.000	0.700	0.879	1.093	1.372	1.812	2.228	2.764	3.169	4.144	4.587
11	0.000	0.697	0.876	1.088	1.363	1.796	2.201	2.718	3.106	4.025	4.437
12	0.000	0.695	0.873	1.083	1.356	1.782	2.179	2.681	3.055	3.930	4.318
13	0.000	0.694	0.870	1.079	1.350	1.771	2.160	2.650	3.012	3.852	4.221
14	0.000	0.692	0.868	1.076	1.345	1.761	2.145	2.624	2.977	3.787	4.140
15	0.000	0.691	0.866	1.074	1.341	1.753	2.131	2.602	2.947	3.733	4.073
16	0.000	0.690	0.865	1.071	1.337	1.746	2.120	2.583	2.921	3.686	4.015
17	0.000	0.689	0.863	1.069	1.333	1.740	2.110	2.567	2.898	3.646	3.965
18	0.000	0.688	0.862	1.067	1.330	1.734	2.101	2.552	2.878	3.610	3.922
19	0.000	0.688	0.861	1.066	1.328	1.729	2.093	2.539	2.861	3.579	3.883
20	0.000	0.687	0.860	1.064	1.325	1.725	2.086	2.528	2.845	3.552	3.850
21	0.000	0.686	0.859	1.063	1.323	1.721	2.080	2.518	2.831	3.527	3.819
22	0.000	0.686	0.858	1.061	1.321	1.717	2.074	2.508	2.819	3.505	3.792
23	0.000	0.685	0.858	1.060	1.319	1.714	2.069	2.500	2.807	3.485	3.768
24	0.000	0.685	0.857	1.059	1.318	1.711	2.064	2.492	2.797	3.467	3.745
25	0.000	0.684	0.856	1.058	1.316	1.708	2.060	2.485	2.787	3.450	3.725
26	0.000	0.684	0.856	1.058	1.315	1.706	2.056	2.479	2.779	3.435	3.707
27	0.000	0.684	0.855	1.057	1.314	1.703	2.052	2.473	2.771	3.421	3.690
28	0.000	0.683	0.855	1.056	1.313	1.701	2.048	2.467	2.763	3.408	3.674
29	0.000	0.683	0.854	1.055	1.311	1.699	2.045	2.462	2.756	3.396	3.659
30	0.000	0.683	0.854	1.055	1.310	1.697	2.042	2.457	2.750	3.385	3.646
40	0.000	0.681	0.851	1.050	1.303	1.684	2.021	2.423	2.704	3.307	3.551
60	0.000	0.679	0.848	1.045	1.296	1.671	2.000	2.390	2.660	3.232	3.460
80	0.000	0.678	0.846	1.043	1.292	1.664	1.990	2.374	2.639	3.195	3.416
100	0.000	0.677	0.845	1.042	1.290	1.660	1.984	2.364	2.626	3.174	3.390
1000	0.000	0.675	0.842	1.037	1.282	1.646	1.962	2.330	2.581	3.098	3.300
<b>Z</b>	0.000	0.674	0.842	1.036	1.282	1.645	1.960	2.326	2.576	3.090	3.291
	0%	50%	60%	70%	80%	90%	95%	98%	99%	99.8%	99.9%
	Confidence Level										