

Terminology, Definitions & Cheat Sheets

STATISTICS

Common Statistics: Various Types of Outcome Data

Outcome Variable	Are the observations independent or correlated?		Alternatives (if assumptions are violated)
	Independent	Correlated / Paired	
Continuous (e.g., profit, price)	Ttest ANOVA Linear correlation Linear regression	Paired ttest Repeated-measures ANOVA Mixed models GEE modeling	Wilcoxon sign-rank test Wilcoxon rank-sum test Kruskal-Wallis test Spearman rank correlation
Binary or categorical (e.g., churn yes/no)	Risk difference Relative risks Chi-square test Logistic regression	McNemar's test Conditional logistic regression GEE modeling	Fisher's exact test McNemar's exact test
Time-to-event (e.g., time to part failure)	Rate ratio Kaplan-Meier statistics Cox regression	Frailty model (beyond the scope)	Time-varying effects (beyond the scope of this course)

Sampling

Term	Definition	Examples/Comments
Bias	A statistical procedure or measure is biased if applied to a sample from a population produces (under-)over-estimates of population characteristic	
Nonresponse Bias	A problem that occurs when non-responders do not show up in surveys	
Parameter	A measurable characteristic of the population	mean, proportion
Population	The target group of study	California voters (eligible to vote? vs. registered?)
Sample	A subset of the population. If drawn randomly, then it is a random sample	
Sampling frame	A practical representation of the population	Only registered voters
SRS: Simple Random Sample	Better known as a randomly drawn sample rather than random sample: each object in the population has an equal chance of being selected	(-) Does not guarantee a fully representative sample (-) Inefficient in practice
Statistic	A measurable characteristic of a sample used to estimate a population parameter	empirical mean is a statistic for a theoretical mean

Sampling Procedure

Term	Definition	Examples/Comments
Bootstrap	A resample with replacement from an observed sample to observe a distribution of a sample statistic	
Randomization	The process of randomly assigning subjects to treatment	
Resample	A collection of observations drawn from the original sample, or generated randomly by a formula based on the original sample	
Sampling with replacement	Each item is replaced after it is drawn from a hat	
Sampling without replacement	Once an item is drawn, it is not eligible to be drawn again: also known as shuffling	
Simulation trial	Taking a resample and calculating the value of some statistic, such as mean, with it	
Simulation	A repeat of multiple single simulation trials and the collection of their calculation results	

Sampling Schemes

Term	Definition	Examples/Comments
Convenience Sampling	There is no effort to define a population or sampling frame: by inviting any one who saw the invite	(+) Easy and cheap (-) Non-representative sample, not well-designed
Cluster Sampling	Clusters of subjects or records selected, and the subjects or records within those clusters are surveyed and measured. Ensure that characteristics that define clusters do not introduce bias into the results	(+) Practical and efficient
Multi-stage Sampling	Randomly select groups and then apply systematic sampling within each group	(+) Minimize cost, sampling error, and bias
Self-Selection	The respondents themselves determine whether they participate in the survey	(-) Biased results
SRS: Simple Random Sample	Better known as a randomly drawn sample rather than random sample: each object in the population has an equal chance of being selected	(-) Does not guarantee a fully representative sample (-) Inefficient in practice
Stratified Sampling	The population is split into categories, or strata , and separate samples are drawn from each stratum.	
Systematic Sampling	Selection of every n^{th} record	

A/B Testing

Term	Definition	Examples/Comments
Control group	A group of subjects exposed to no (or standard) treatment	
Randomization	The process of randomly assigning subjects to treatment	
Standard Error	The variability (standard deviation) of a sample statistic over many samples	Unlike standard deviation that refers to variability of values within a single sample
Subjects	The items that are exposed to treatments	web visitors, patients, ..
Test statistic	The metric used to measure the effect of treatment	t-statistic
Treatment	Something to which a subject is exposed	drug, price, web headline
Treatment group	A group of subjects exposed to a specific treatment	

Confidence Interval (CI)

Term	Definition	Examples/Comments
Central Limit Theorem (CLT)	CLT says that the means drawn from multiple samples will be Normally distributed, even if the population is not Normally distributed if the sample size is “large enough” (20-30) and the departure from Normality is “not great”	Allows Normal-approximation formulas to be used in calculating sampling distributions for CIs and hypothesis tests
Confidence Interval	A 90% confidence interval for the mean (or other statistic) is a range that encloses the central 90% of the resampled means (or other statistic). 90% CI encloses the true statistic 90% of the time when constructed repeatedly in the same manner with the same population.	
Margin of Error	A plus or minus quantity attached to a point estimate, whereas a CI is the actual endpoints of the interval	If the mean of the positive rate is 36% and CI is [32,40], then margin of error is 36% +/-4%
Point Estimate	A statistic calculated from a sample	mean (\bar{x}), median, quantile
Standard Error	Standard error (s.e.) of a sample statistic is the standard deviation of the sample statistic (s.e. of the estimate)	sample estimate means sample statistic

Hypothesis Tests

Term	Definition	Examples/Comments
Alternate Hypothesis	The theory you would like to accept, assuming that your results disprove the null hypothesis	Counterpoint to the null (what you hope to prove)
Alpha (α) Level or Significance	The threshold level of statistical significance	Determined before a study is done: e.g. 5% is customary
Hypothesis Test, or Significance Test	Answers the questions, “Might this apparently interesting result have happened by chance owing to the luck of the draw in who/what gets selected in sample(s) or assigned to difference treatments?”	CI: “How much chance error might be in this measurement or estimate or model, owing to the luck ...?”
Null Hypothesis Null Model	An imaginary chance model representing the idea that nothing new is going on. The hypothesis that chance is to blame.	It nullifies what you are trying to prove (e.g., no difference between treatments A and B)
<i>p</i> -value	The frequency with which a result as extreme as observed results occurs just by chance, drawing from the null hypothesis model	$p\text{-value} \leq 0.05$ indicates statistical significance → rejects the Null Hypothesis. It should be low enough to meet the threshold established (alpha) for statistical significance
Statistical Inference	The process of accounting for random variation in data as you draw conclusions	

Hypothesis Tests (cont.)

Term	Definition	Examples/Comments
One-tailed test or One-way test	The whole alpha α is allotted to only one direction of the distribution	Hypothesis test that counts chance results only in one direction
Test statistic	The metric used to measure the effect of treatment	t-statistic, F-statistic, χ^2 -statistic
Two-tailed test or Two-way test	half of alpha α is used to test the statistical significance in one direction of the distribution and the other half in the other direction	Hypothesis test that counts chance results only in two direction
Type I Error	Reject H_0 when, in fact, H_0 is true. Mis-takingly concluding an effect is real (when it is due to chance)	a jury convicted an innocent person for a crime that the person did not commit
Type II Error	Fail to Reject H_0 when, in fact, H_1 is true. Mis-takingly concluding an effect is due to chance (when it is real)	a guilty person escaped conviction

Statistical Distributions & Functions in R

Distribution	Random Number Generator	Density	Distribution	Quantile
Normal	r norm	d norm	p norm	q norm
t	rt	dt	pt	qt
F	rf	df	pf	qf
χ^2	rchisq	dchisq	pchisq	qchisq

{dpr}distribution_abbreviation()

- **d** = density
 - **p** = distribution function
 - **q** = quantile function
 - **r** = random generation
-
- **pnorm(a)** $\equiv P(X \leq a)$: probability that a or smaller number occurs
 - **pnorm(b) – pnorm(a)** $\equiv P(a \leq X \leq b)$: probability that the variable falls between two points
 - **qnorm()**: given the cumulative probability distribution, it returns the quantile

Statistical Distributions: Mean and Variance

Distribution	Degrees of freedom	Mean	Variance
Normal		μ	σ^2
t	n	0	$n/(n - 2)$
F	n_1 and n_2	$n_2/(n_2 - 2)$	a/b
χ^2	r	r	$2r$

$$a = 2n_2^2(n_1 + n_2 - 2)$$

$$b = n_1(n_2 - 2)^2 (n_2 - 4)$$

Proxy Statistic: Hypothesis Testing & Confidence Intervals

Aim	Model Statistic	Sample Statistic	Proxy Statistic	Formula for Proxy
Estimate the mean μ of a normal distribution with known variance σ^2	μ	m	Z-statistic	$Z \sim \frac{m - \mu}{\sigma / \sqrt{n}}$
Estimate the variance σ^2 of a normal distribution with known mean μ	σ^2	S^2	χ^2 -statistic	$\chi^2_{n-1} \sim (n-1) \frac{S^2}{\sigma^2}$
Estimate the mean μ of a normal distribution with un-known variance σ^2	μ	m	t-statistic	$T_{n-1} \sim \frac{m - \mu}{S / \sqrt{n}}$

Ex.	Proxy Statistic	Distribution	Degrees of Freedom (df)
1	Z-statistic	$N(0, 1)$	
2	χ^2 -statistic	$\chi^2(n-1)$	$n-1$
3	t-statistic	T_{n-1}	$n-1$

Hypothesis Testing: Procedure

- Step 1: Define **a statistic** that obeys a certain **distribution** if the hypothesis is correct:
 - Ex-1: The mean μ from a normal distribution with known variance σ^2
 - Ex-2: The variance σ^2 from a normal distribution with known mean μ
 - Ex-3: The mean μ from a normal distribution with unknown variance σ^2
- Step 2 (optional): Transform the statistic to a **proxy statistic** with the **proxy distribution** of better understood properties/characteristics:
 - Ex-1: Z-statistic from a uniform normal distribution, $N(0,1)$
 - Ex-2: χ_{n-1}^2 -statistic from a χ^2 distribution with n df
 - Ex-3: T_{n-1} -statistic from a t -distribution with $n - 1$ df
- Step 3: Calculate the statistic (original/proxy) from the **sample**
- Step 4: Compute the **probability** (the **p-value**) of this sample with this statistic to be drawn from this distribution (original/proxy)
 - **Reject the hypothesis** if probability is **low** (e.g., **p-value < 0.05**)
 - **Fail to reject the hypothesis** otherwise (e.g., **p-value \geq 0.05**)

Hypothesis Tests

Sample	Paired	Null Hypothesis	Assumptions	Test
One Sample		$H_0 : \mu = \mu_0$	i.i.d. $N(\mu, \sigma^2)$	t.test()
Two Samples	No	$H_0 : \sigma_1^2 = \sigma_2^2$	Normally distributed	F-test: var.test() Bartlett: bartlett.test()
Two Samples	No	$H_0 : \sigma_1^2 = \sigma_2^2$	Non-parametric	Ansari-Bradley test: ansari.test()
Two Samples	No	$H_0 : \mu_1 = \mu_2$	$\sigma_1^2 = \sigma_2^2$	t.test(var.equal=TRUE)
Two Samples	No	$H_0 : \mu_1 = \mu_2$	$\sigma_1^2 \neq \sigma_2^2$	Welch t-test t.test(var.equal=FALSE)
Two samples	No	$p_1(x) = p_2(x)$ p : probab. distr	Non-parametric	Wilcoxon rank sum wilcox.test ()
Two Samples	Yes	$H_0 : \mu_1 = \mu_2$	$\sigma_1^2 \neq \sigma_2^2$	t.test(paired=TRUE)
Two samples	Yes	$p_1(x) = p_2(x)$ p : probab. distr	Non-parametric	wilcox.test (paired=TRUE)

Power and Sample Size

Term	Definition	Examples/Comments
Effect size	The minimum size of the effect that you hope to be able to detect in a statistical test	20% improvement in click rates
Statistical Power	The probability of detecting a given effect size with a given sample size	
Significance level	The statistical significance level at which the test will be conducted	
Type I Error	Reject H_0 when, in fact, H_0 is true. Mis-takingly concluding an effect is real (when it is due to chance)	a jury convicted an innocent person for a crime that the person did not commit
Type II Error	Fail to Reject H_0 when, in fact, H_1 is true. Mis-takingly concluding an effect is due to chance (when it is real)	a guilty person escaped conviction

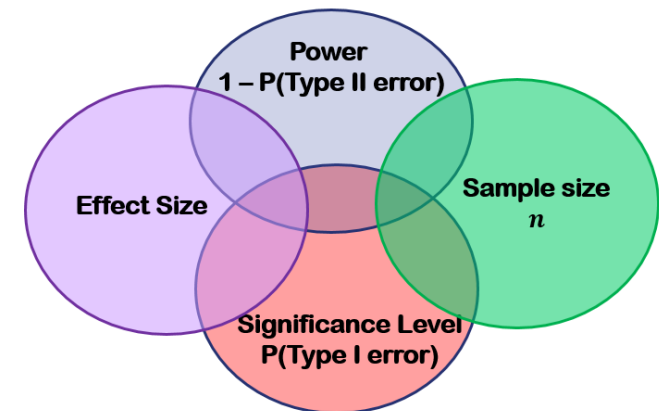
Type I and II Errors, Power, Significance Level (α)

Actual	Decision	
	Reject H0	Fail to Reject H0
	H0 True	H0 False
	Type I Error	correct
	correct	Type II Error

Power = 1 – Probability (Type II Error)

α = Probability (Type I Error)

- **Type I Error**
 - Probability of False Positives
 - Fail to find an effect that IS there
- **Type II Error**
 - The probability of False Negatives
 - Find an effect that is NOT there
- **Power** is defined as one minus the probability of making a Type II error
 - The probability of finding an effect that *IS* there
- **Significance level (α)** is the probability of making a Type I error
 - The probability of finding an effect that is *NOT* there



Multiple Testing

Term	Definition	Examples/Comments
Type I Error	Mistakenly concluding that an effect is statistically significant	
False Discovery Rate (FDR)	Across multiple tests, the rate of making Type I error	
Adjustment of p -value	Accounting for doing multiple tests on the same data	
Overfitting	Fitting the noise	

Multiple Testing: Sources & Mitigation

- Sources of **multiplicity** issues:
 - Multiple comparisons in multiple tests of significance
 - Many variables
 - Many models
- Multiplicity increases the risk of concluding that something is significant just by chance (**Type I error**)
- Mitigation strategies for multiple statistical comparisons
 - **Adjustment procedure**: dividing alpha by the number of tests
 - **Bonferroni adjustment**: dividing alpha by the number of observations, n
- Mitigation strategy for supervised modeling
 - **Cross-validation**: holdout sample with labeled outcome variables

Design #1: One-way **Between-Groups** ANOVA

- Exemplar Study: Goal: To study two treatments of anxiety
 - **Treatment (Independent Variable):** Two treatments
 - CBT: Cognitive Behavior Therapy
 - EMDR: Eye Movement Desensitization & Reprocessing Therapy
 - **Response (Dependent Variable):** (collected after 5 weeks of treatment)
 - STAI: State-Trait Anxiety Inventory; a self-report measure of anxiety
 - **Subjects:**
 - Randomly divided between two **independent groups**: CBT & EMDR

		Subjects
Treatment	CBT	s_1
		s_2
		...
		$s_{n/2}$
	EMDR	$s_{n/2+1}$
		...
		...
		s_n

One-way Between-Groups **Balanced** ANOVA

- Treatment is a **between-groups** factor with two levels
- **Balanced design**: equal number of subjects in each treatment condition; otherwise, **unbalanced**
- **One-way**: because a single classification variable

F-tests to assess the effects in ANOVA designs

- If the F-test for Treatment is significant then reject the null hypothesis: H_0 : the mean STAI scores are the same
- Conclude: the mean STAI scores changed over time for two therapies differed after five weeks of treatment

Design #2: One-way **Within-Groups** ANOVA

- Exemplar Study: Goal: To study **longitudinally** one treatment of anxiety
 - **Treatment (Independent Variable):** One treatment
 - CBT: Cognitive Behavior Therapy
 - **Response (Dependent Variable):** (collected after 5 weeks of treatment)
 - STAI: State-Trait Anxiety Inventory; a self-report measure of anxiety
 - **Subjects:**
 - The same subjects over different time points (**dependent groups**)
 - **Time:** Two different time points: 5 weeks & 6 months

One-way Within-Groups Balanced ANOVA

- Time is a **within-groups** factor with two levels: each subject is measured under both levels
 - **Repeated measures** ANOVA
- **One-way:** because a single classification variable

Paired F-tests to assess effects in ANOVA designs

- If the F-test for Treatment is significant then reject the null hypothesis: H_0 : the mean STAI scores are the same
- Conclude: the mean STAI scores change over time: between 5 weeks and 6 months

	Time	
	5 weeks	6 months
CBT Treatment	s_1	s_1
	s_2	s_2

	s_n	s_n

Design #3: Factorial (Mixed-Model) ANOVA

- Exemplar Study: Goal: To study **two treatments** of anxiety **longitudinally**
 - **Treatment (Independent Variable):** Two treatments
 - CBT: Cognitive Behavior Therapy
 - EMDR: Eye Movement Desensitization & Reprocessing Therapy
 - **Response (Dependent Variable):** (collected after 5 weeks of treatment)
 - STAI: State-Trait Anxiety Inventory; a self-report measure of anxiety
 - **Subjects:** Randomly assigned to two **independent groups**: CBT & EMDR
 - **Time:** Two **dependent groups** over time: 5 weeks & 6 months

		5 wks	6 mo.
Treatment	CBT	s_1	s_1
		s_2	s_2
	
		$s_{n/2}$	$s_{n/2}$
		$s_{n/2+1}$	$s_{n/2+1}$
	EMDR
	
	
		s_n	s_n

Two-way Factorial ANOVA Design

- **Main effects:** Impact of Therapy (averaged across Time) and Time (averaged across Therapy type)
- **Interaction effect:** Interaction of Therapy & Time
- **Factorial ANOVA:** cross 2 factors (**two-way**) or more

Three F-tests to assess ANOVA design effects

- F-test for Therapy: Significant → CBT & EMDR differ
- F-test for Time: Significant → change over time
- F-test for Therapy × Time interaction: Significant → two treatments had a differential impact over time: different change from 5 wks to 6 mo. for 2 therapies

Design #4: ANCOVA

- Goal: To study the treatments of anxiety with **confounding factors**
- **Confounding Factor (Nuisance Variable)**: Other factors (than Treatment) that could explain the post-therapy differences on the dependent variable
 - Depression level:
 - A self-reported measure such as Beck Depression Inventory (BDI)
 - Although subjects were assigned randomly to treatment conditions, it is possible that two therapy groups differed in patient depression levels at the start of the study
 - → Any post-therapy differences might be due to the preexisting depression differences and not to experimental manipulations

ANCOVA Design

- Because depression could also explain the group difference on the dependent variable, it is a **confounding factor**
- Because the study is not interested in depression, this confounding factor is called a **nuisance variable**

Design #5-6: **MANOVA** & **MANCOVA**

- Goal: To study the treatments of anxiety with **multiple dependent variables** and/or **confounding factors**
- **Multiple Dependent Variables:** To increase the validity of the study
 - STAI: One dependent variable
 - Family ratings: Another dependent variable
 - Therapy ratings: Yet another dependent variable
 - A measure assessing the impact of anxiety on the daily functioning
- **MANOVA: Multivariate Analysis of Variance**
 - There is more than one dependent variable
- **MANCOVA: Multivariate Analysis of Covariance**
 - Besides multiple dependent variables, there are covariates present