STAT 135: Concepts of Statistics

—Final Review Guide—

Disclaimer: This study guide was created by the Statistics Undergraduate Student Association, which is in no way affiliated with the current course staff of Statistics 135. Additionally, this guide is designed merely to *supplement* to your class notes.

Last Updated: Fall 2020

Originally Compiled by Ethan Marzban

Contents

1 Inferential Statistics

- A **population** is too large to be observed in its entirety; as such, we must use **samples** to explore properties of the population.
  - The gold-standard of sampling is a **Simple Random Sample**, in which observations are taken independently of each other, from identical distributions.
  - Other sampling techniques include **stratified sampling**, and **cluster sampling.**
- **Population parameters** are deterministic quantities pertaining to the population; their exact values can never be determined exactly, and must be estimated using **estimators** (which are functions of data).
  - **Bias** measures "how far off" an estimator is from the parameter it is estimating. Mathematically, if $\hat{\theta}$ is an estimator for θ, we write $\text{Bias}(\hat{\theta}, \theta) = \mathbb{E}(\hat{\theta}) - \theta$.
- **Confidence Intervals** provide a way of quantifying uncertainty, specifically with respect to estimations of population parameters.
- The **Central Limit Theorem** provides information on the asymptotic behavior of the sample mean.
- Samples are generally taken *without* replacement, and observations are therefore (technically) dependent. However, as the sample size increases, the dependence between observations tends towards 0.
  - The **finite population correction factor** relates samples drawn without replacement to those drawn with replacement. When the sample size is very large, the finite population correction factor is approximately 1.

2 Parameter Estimation

- Two popular estimators of population parameters are the **Method of Moments (MoM)** estimator and the **Maximum Likelihood Estimator (MLE)**.
  - MLE's satisfy the **equivariance property** (sometimes called the **invariance property**), which states that $\widehat{f(\theta)}_{\text{ML}} = f(\hat{\theta}_{\text{ML}})$ (provided f satisfies certain mathematical properties).
  - The variance of the MoM estimator can be approximated using the **method of propagation of errors** (also known as the **δ-method**).
  - The asymptotic variance of the MLE is $[I_n(\theta)]^{-1}$, where $I_n(\theta)$ is the **Fisher information** of the sample.
- The **Cramér-Rao Lower Bound (CRLB)** provides a lower bound on the variance of *any* unbiased estimator of a population parameter (under certain regularity conditions).
  - Estimators whose variance are exactly equal to the CRLB are said to be **efficient**.
- **Sufficient Statistics** are functions (of data) that contain all the information about a parameter θ, given a sample. Mathematically, they induce a partition of the sample space that is finer than (or as fine as) the likelihood function. A **Minimal Sufficient Statistic** is a sufficient statistic that partitions the sample space in the *coarsest* manner.
  - Sufficient statistics are not unique, whereas minimally sufficient statistics are.[1]

[1] Technically, if T and U are both minimally sufficient statistics then there exists a one-to-one function ϕ such that $T = \phi(U)$, so it would be more mathematically rigorous to say $T \sim U$ rather than $T = U$.

- Sufficient statistics are typically found using the **Factorization Theorem**.

- An estimator $\hat{\theta}_n$ of θ is said to **converge in probability** to θ if $\mathbb{P}(|\hat{\theta}_n - \theta| \geq \varepsilon) \to 0$ as $n \to \infty$. Convergence in probability is notated as: $\hat{\theta}_n \xrightarrow{\text{p}} \theta$.
  - Compare this with **convergence in distribution:** a sequence $\{X_i\}_{i=1}^n$ is said to **converge in distribution** to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

    Convergence in distribution is *weaker* than convergence in probability.

- An estimator $\hat{\theta}$ is said to be a **consistent** estimator of a parameter θ if $\hat{\theta} \xrightarrow{\text{p}} \theta$.
  - As an example: under certain regularity conditions, the MLE is consistent.

- Given an estimator $\hat{\theta}$ of a parameter θ, a "better" (i.e. lower-variance) estimator can always be obtained by conditioning on a sufficient statistic. This is known as **Rao-Blackwellization**.

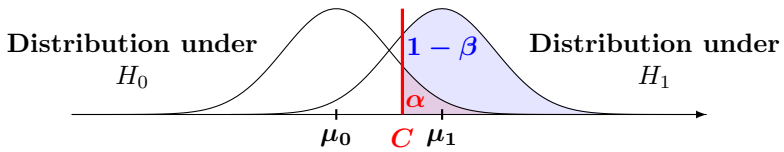
- The **bootstrap** provides another method for parameter estimation.
  - In the **nonparametric bootstrap**, no assumptions are made about the underlying distribution. The sampling distribution is approximated by repeatedly sampling (with replacement) from the original sample, and the remainder of inference is conducted as before.
  - In the **parametric bootstrap**, assumptions are made about the underlying distribution. The parameters of said distribution are computed from the original sample, and the sampling distribution is approximated by repeatedly generating samples from the assumed distribution (with the estimated parameters plugged in).

3 Hypothesis Testing

- The **null hypothesis** is chosen to represent the status quo; the **alternative hypothesis** provides a theory contrary to the null hypothesis. The goal of **hypothesis testing** is to determine which of the two hypotheses better describes the current state.

- There are several terms and notations associated with hypothesis testing:
  - **Level of Significance** (α): $\mathbb{P}(\text{reject } H_0 \mid H_0 \text{ is true})$
    - The event $\{(\text{reject } H_0 \mid H_0 \text{ is true}\}$ is known as a **Type I Error**
  - **Power [of a test]** $(1-\beta)$: $\mathbb{P}(\text{fail to reject } H_0 \mid H_0 \text{ is false})$
    - The event $\{(\text{fail to reject } H_0 \mid H_0 \text{ is false}\}$ is known as a **Type II Error**

Result of Test	State of H_0: **True**	**False**
Reject	**Type I**	No Error
Fail to Reject	No Error	**Type II**



- The **Neyman-Pearson Lemma** states that the **Generalized Likelihood Test** is **uniformly most powerful**.

- Hypothesis testing and confidence intervals are equivalent.

- $p-$values are always constructed "under the null;" that is, they are computed *after* assuming the null hypothesis is true.
- The **rejection region** is the set of values of a test statistic that lead to rejection of the null: $\mathcal{R} = \{x_i : \text{test at } x_i \text{ rejects } H_0\}$.

4 Partitioning the Sample Space

- A **partition** can be thought of a grouping structure that groups like elements together. For example, consider the likelihood function of our scenario above:

$$p(\theta) = \prod_{i=1}^{3} p(x_i;\theta) = \prod_{i=1}^{3} \theta^{x_i}(1-\theta)^{1-x_i} = \theta^{\sum_{i=1}^{3} X_i}(1-\theta)^{3-\sum_{i=1}^{3} X_i}$$

- There are only **8** possible configurations of our data (X_1, X_2, X_3); let us enumerate them all:

$$\{(0,0,0),(0,0,1),(0,1,0),(1,0,0),(0,1,1),(1,0,1),(1,1,0),(1,1,1)\}$$

  The set of all possible configurations of data is called the **sample space**. So, the set above would be the sample space of our example.

- Each of these **8** possibilities corresponds to a different value of T:

(X_1, X_2, X_3)	$(0,0,0)$	$(0,0,1)$	$(0,1,0)$	$(1,0,0)$	$(0,1,1)$	$(1,0,1)$	$(1,1,0)$	$(1,1,1)$
$T(X_1, X_2, X_3)$	0	1	1	1	2	2	2	3

- Here, we can see how the statistic T induces a partition: it is natural to group all $T = 0$ terms together, all $T = 1$ terms together, and all $T = 2$ terms together.
- More generally, a partition is a collection of sets $\{B_1, \ldots, B_n\}$. A partition is defined to be **sufficient** if $f(x_i \mid X_1, \ldots, X_n \in B)$ does not depend on θ. By "the partition induced by T", we mean the partition consisting of elements $\{t : T(X_1, X_2, X_3) = t\}$ for all possible values of t. Therefore, for a statistic to be sufficient, the partition it induces must be sufficient.
- Algorithmically, to determine whether or not a *partition* is sufficient, we compare it to the partition induced by the conditional probability $\mathbb{P}(X_1, X_2, X_3 = (x_1, x_2, x_3) \mid T)$. For example,

$$p[(0,0,1) \mid T = 1] = \frac{\mathbb{P}(X_1 = 0, X_2 = 0, X_3 = 0 \mid T = 1)}{\mathbb{P}(T = 1)} = \frac{\theta(1-\theta)^2}{\theta(1-\theta)^2 + \theta(1-\theta)^2 + \theta(1-\theta)^2} = \frac{1}{3}$$

  Iterating through the **8** possibilities, we obtain the following table:

(X_1, X_2, X_3)	$T(X_1, X_2, X_3)$	$p(X_1, X_2, X_3)$
(0, 0, 0)	0	$\frac{(1-\theta)^3}{(1-\theta)^3} = 1$
(0, 0, 1)	1	$1/3$
(0, 1, 0)	1	$1/3$
(1, 0, 0)	1	$1/3$
(0, 1, 1)	2	$1/3$
(1, 0, 1)	2	$1/3$
(1, 1, 0)	2	$1/3$
(1, 1, 1)	3	1

  Clearly the partition induced by $T = \sum X_i$ does not depend on θ; hence T is sufficient.

- As an example of a **non-sufficient** (or insufficient) statistic for θ, consider $T_2 = X_1$. Let us also examine one of the conditional probabilities in more detail:

$$p\left[(0,0,0) \mid T=0\right] = \frac{\mathbb{P}(X_1=0, X_2=0, X_3=0 \mid T=0)}{\mathbb{P}(T=0)} = (1-\theta)^2$$

Here is the full partition induced by this statistic:

(X_1, X_2, X_3)	$T_2(X_1, X_2, X_3)$	$p(X_1, X_2, X_3)$
(0, 0, 0)	0	$(1-\theta)^2$
(0, 0, 1)	0	$\theta(1-\theta)$
(0, 1, 0)	0	$\theta(1-\theta)$
(1, 0, 0)	1	$\theta(1-\theta)^2$
(0, 1, 1)	0	θ^2
(1, 0, 1)	1	$\theta(1-\theta)$
(1, 1, 0)	1	$\theta(1-\theta)$
(1, 1, 1)	1	θ^2

- Note that the **likelihood** itself induces a partition:

(X_1, X_2, X_3)	$L(\theta; X_1, X_2, X_3)$
(0, 0, 0)	$(1-\theta)^3$
(0, 0, 1)	$\theta(1-\theta)^2$
(0, 1, 0)	$\theta(1-\theta)^2$
(1, 0, 0)	$\theta(1-\theta)^2$
(0, 1, 1)	$\theta^2(1-\theta)$
(1, 0, 1)	$\theta^2(1-\theta)$
(1, 1, 0)	$\theta^2(1-\theta)$
(1, 1, 1)	θ^3

This so-called **likelihood partition** provides a good test for whether or not statistics are sufficient. If the likelihood partition has divisions (denoted by horizontal lines in the tables above) if is not sufficient. Otherwise, it is sufficient:

(X_1, X_2, X_3)	$T(X_1, X_2, X_3)$	$T_2(X_1, X_2, X_3)$	$L(\theta; X_1, X_2, X_3)$
(0, 0, 0)	0	0	$(1-\theta)^3$
(0, 0, 1)	1	0	$\theta(1-\theta)^2$
(0, 1, 0)	1	0	$\theta(1-\theta)^2$
(1, 0, 0)	1	1	$\theta(1-\theta)^2$
(0, 1, 1)	2	0	$\theta^2(1-\theta)$
(1, 0, 1)	2	1	$\theta^2(1-\theta)$
(1, 1, 0)	2	1	$\theta^2(1-\theta)$
(1, 1, 1)	3	1	θ^3

Note that the partition induced by T_2 "broke" a horizontal line; as such, it is not sufficient.

- A **minimal sufficient partition** is the coarsest sufficient partition. It can be shown that the likelihood always generates the coarsest sufficient partition; hence, we adopt the following "test:" a statistic is minimal sufficient (for θ) if the partition it induces is the same as the likelihood partition.

- In our example above, T_1 generates the same partition as the likelihood and is therefore **minimal sufficient**.

Rule-of-Thumb: If the likelihood partition creates divisions where the partition induced by T does not, then T is **not sufficient**; otherwise it is **sufficient**. If it T sufficient, and partitions the sample

space in the same way as the likelihood, then it is **minimal sufficient**.

5 Two-Sample Tests

- **Big Idea:** given two samples $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, we wish to determine whether or not these samples came from the same distribution.
  - This question, however, is too broad. Instead, we will compare **sample means**. That is, we test the hypothesis that $\mu_X = \mu_Y$, where μ_X and μ_Y denote the populations means of the distributions from which X and Y were sampled, respectively.
- **Key Assumption:** we will assume that the two samples come from **independent normal distributions**. That is,
$$X \sim \mathcal{N}(\mu_X, \sigma_X^2); \qquad Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$
In this framework, our null hypothesis is
$$H_0 : \ \mu_X = \mu_Y$$
  - There are several choices for the alternative hypothesis: for example, $\mu_X \neq \mu_Y$ (a **two-sided alternative**), or various **one-sided alternatives** (e.g. $\mu_X > \mu_Y$, or $\mu_X < \mu_Y$).
- The statistic we use will always be of the form
$$T = \frac{(\bar{X} - \bar{Y}) - (\mu_X - \mu_Y)}{\hat{\sigma}\sqrt{\frac{1}{n} + \frac{1}{m}}}$$
where $\hat{\sigma}$ denotes an estimate of standard deviation. There are two possible estimators we can use to estimate σ, depending on whether or not we assume $\sigma_X = \sigma_Y$.
  - If we assume $\sigma_X = \sigma_Y := \sigma$ (that is, if we assume both X and Y have the same population standard deviation), use
$$\hat{\sigma}^2 = \frac{(n-1)s_X^2 + (m-1)s_Y^2}{m+n-2}$$
(called the **pooled sample variance**). With this estimate, $T \sim t_{m+n-2}$.
  - If we assume $\sigma_X \neq \sigma_Y$ (that is, if we assume X and Y have the *different* population standard deviations), use
$$\hat{\sigma}^2 = \frac{s_X^2}{n} + \frac{s_Y^2}{m}$$
(where s_X^2 and s_Y^2 denote the sample standard deviations from $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^n$, respectively). With this estimate, **the distribution of T is unknown**, however it is (asymptotically) very close to a t distribution with degrees of freedom
$$\text{df} = \text{round}\left\{ \frac{\left[\left(\frac{s_X^2}{n}\right) + \left(\frac{s_Y^2}{m}\right)\right]^2}{\frac{\left(\frac{s_X^2}{n}\right)^2}{n-1} + \frac{\left(\frac{s_Y^2}{m}\right)^2}{m-1}} \right\}$$
- Everything above outlines a set of *parametric* tests; we can also use a very popular *nonparametric* test, known as the **Mann-Whitney Test** (also known as the **Wilcoxin Rank-Sum Test**).
  - The main idea behind the Mann-Whitney test is as follows. Suppose we have two samples, $\{X_i\}_{i=1}^n$ and $\{Y_i\}_{i=1}^m$, and we wish to ascertain whether or not these samples came from the same distribution. We therefore set this to be our null: H_0 asserts that the samples were drawn from the same population.

- We now assume the null hypothesis (as we do for all statistical testing). Since we believe both samples to have come from the same population, let's pool our sample observations together to form a set of $N := n + m$ observations.

  If we see that the majority of observations from X are smaller (or larger) than those from Y, we have reason to believe that there *is* some strong underlying difference between X and Y, and we would reject the null.

  On the other hand, if we see that all values from X and Y are mixed together well, with no noticeable differences, we would have no reason to reject the null.
- To actually carry out this comparison, we assign **ranks** to the N pooled observations, and examine the possible distribution of ranks. This forms the basis of the Mann-Whitney Test.

- **Algorithm: Mann-Whitney Test**

  (1) Pool the $N := n+m$ samples together, and assign N ranks to these pooled observations. (If there are ties, assign each the average of the tied ranks).

  (2) Let T_Y denote the sum of a random sample of size m from these ranks. Find the distribution of T_Y:
  - If the sample size is small, enumerate all possible permutations by hand.
  - If the sample size is large, you can approximate the distribution of T_Y with a normal distribution.

  (3) Return to the original configuration of your data, and determine the "sample rank-sum" (that is, the sum of ranks in the actual sample $\{Y_i\}_{i=1}^m$) and place this on the distribution of T_Y.

  (4) Form a rejection region based on a level of significance α and your statistic from step (3) above.

6 Analysis of Variance (ANOVA)

- In a **one-way layout**, we consider I groups, each with some number of measurements on *one* particular factor, and attempt to explain the difference in means across groups. (Compare this with a **two-way layout**, in which two factors are measured within each group).
  - **Example of a one-way layout:** Consider scores on an exam, grouped by year (i.e. Freshman, Sophomore, Junior, Senior), and suppose we want to determine whether or not there is a significant difference in the average test score for the different years.
  - Two-way layouts are generally not covered in Stat 135.

- **ANOVA** (Analysis of Variance) is a type of Hypothesis Testing; the null and alternate hypotheses are as follows:

$$\begin{cases} H_0 : \text{Differences in means are due purely to chance} \\ H_1 : \text{Differences in means are due to some confounding variable} \end{cases}$$

- **The Model:** Letting Y_{ij} denote the jth observation of the ith group, and letting α_i denote the differential effect of the ith treatment, we can write our model as

$$Y_{ij} = \mu + \alpha_i + \varepsilon_{ij}$$

where we impose the condition $\sum_{i=1}^{I} \alpha_i = 0$. **Key Assumption:** $\varepsilon_{ij} \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, \sigma^2)$. [This assumption can be relaxed in a nonparametric setting; see the notes on the Kruskal-Wallis Test below].

- $\text{SS}_\text{T} = \text{SS}_\text{W} + \text{SS}_\text{B}$. That is: Sum of Squares, Total is equal to the Sum of Squares Within [each group] plus the Sum of Squares Between [groups]. For now, assume each group has the same number of observations (which we call J); then, mathematically, we write

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y})^2}_{\text{SS}_\text{T}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y}_{i\cdot})^2}_{\text{SS}_\text{W}} + \underbrace{J\sum_{i=1}^{I}(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2}_{\text{SS}_\text{B}} \qquad (\text{SS}_\text{T}\text{ Decomposition})$$

  where

$$\bar{Y}_{i\cdot} := \frac{1}{J}\sum_{j=1}^{J} Y_{ij}; \qquad \bar{Y}_{\cdot\cdot} := \frac{1}{IJ}\sum_{i=1}^{I}\sum_{j=1}^{J} Y_{ij}$$

- Another way to phrase the null hypothesis, in terms of the α_i's, is

$$H_0 : \alpha_1 = \cdots = \alpha_I = 0$$

  To test this hypothesis, we use

$$F = \frac{\text{SS}_\text{B}/(I-1)}{\text{SS}_\text{W}/[I(J-1)]}$$

  which, under the null, follows an $F_{I-1,\ I(J-1)}$. We reject H_0 for large values of F.

- **A Key Problem:** [This appeared as **Problem 1** on the Midterm Review Problems sheet, curated by SUSA.] Suppose H_0 is true, and suppose we perform n hypothesis tests on H_0, each at an α level of significance. Then, on average, we expect to reject $n\alpha$ hypotheses *even though they are all true.*
  - To mitigate this, we use the **Bonferroni Correction**, which states that each hypothesis test should be conducted at an α/n level of significance. This ensures the overall significance level is less than or equal to α.
  - Often times, when performing ANOVA, we adopt a Bonferroni Correction, dividing α by the number of pairwise comparisons possible. Thus, if there are J groups, we divide by $\binom{J}{2}$.
  - **Alternate Formulation of ANOVA:** Construct confidence intervals for the true mean of observations within each group, using a Bonferroni Correction on the level of significance α. If all J of these confidence intervals overlap, fail to reject the null. Otherwise; reject the null (i.e., conclude that at least one group has a significantly different mean than the others).

- **CRUCIAL NUANCE:** The alternative hypothesis in ANOVA is NOT: "all groups have different means." Rather, the alternative is the logical negation of the statement "all groups have the same mean," which is

$$H_A : \text{At least one group has a significantly different mean than the others}$$

- **Unequal Group Sizes:** Suppose that we still have I groups, but that each group does *not* have the same number of observations. In this case, let J_i denote the number of observations in group $i \in \{1, \ldots, I\}$; the SS_T Decomposition can be modified as

$$\underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J}(Y_{ij}-\bar{Y})^2}_{\text{SS}_\text{T}} = \underbrace{\sum_{i=1}^{I}\sum_{j=1}^{J_i}(Y_{ij}-\bar{Y}_{i\cdot})^2}_{\text{SS}_\text{W}} + \underbrace{\sum_{i=1}^{I} J_i(\bar{Y}_{i\cdot}-\bar{Y}_{\cdot\cdot})^2}_{\text{SS}_\text{B}} \qquad (\text{SS}_\text{T}\text{ Decomposition v2})$$

  where

$$\bar{Y}_{i\cdot} := \frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}; \qquad \bar{Y}_{\cdot\cdot} := \frac{1}{I}\sum_{i=1}^{I}\bar{Y}_{i\cdot} = \frac{1}{I}\sum_{i=1}^{I}\frac{1}{J_i}\sum_{j=1}^{J_i} Y_{ij}$$

- **Alternative Formulation of SS_W:** Let s_i^2 denote the sample variance of observations in group I. Then

$$\text{SS}_\text{W} = \sum_{i=1}^{I} (J_i - 1) s_i^2$$

If we again assume the sample sizes across groups are the same, and equal to J, then an unbiased estimate of σ^2 (the true population variance of the errors) is given by

$$s_p^2 := \frac{\text{SS}_\text{W}}{I(J-1)}$$

- **What Formula to use When?** We've seen two forms of SS_W; one involving Y_{ij}'s, and the other involving s_i^2's. The reason we might chose to use the second formula over the first is if we don't have access to the original data! In some problems, you will be given *only* sample sizes and sample variances, and will *not* be given the original data. Clearly, in this case, you *can't* use the first form of SS_W!

- If the assumption of Gaussian errors is not met, a nonparametric analog of ANOVA can be used. [Independence across observations is still assumed.] One such test is the **Kruskal-Wallis Test**, which can be thought of as a generalization of the Mann-Whitney test. Unsurprisingly, it utilizes ranks.

- **Algorithm: Kruskal-Wallis Test.** H_0 : Observations within each group have the same distribution.

  (1) Combine all observations Y_{ij} into a single set. Define

$$R_{ij} = \text{Rank of } Y_{ij} \text{ in the combines sample}$$

$$\overline{R}_{i\cdot} = \text{Average of ranks of observations in group } I = \frac{1}{J_i} \sum_{i=1}^{J_i} R_{ij}$$

$$\overline{R}_{\cdot\cdot} = \text{Global average of Ranks} = \frac{1}{N} \sum_{i=1}^{I} \sum_{j=1}^{J_i} R_{ij} = \frac{1}{N} \cdot \frac{N(N+1)}{2} = \frac{N+1}{2}$$

  where $N := \sum J_i$ denotes the total number of observations.

  (2) Compute SS_B, where

$$\text{SS}_\text{B} = \sum_{i=1}^{I} J_i (\overline{R}_{i\cdot} - \overline{R}_{\cdot\cdot})^2$$

  (this can be thought of as a measure of how spread out the ranks are from the average rank). Under H_0, SS_B should be small.

  (3) Compute the sample variance of the R_{ij}'s, using

$$\text{Var}(R) = \frac{1}{N-1} \sum_{i=1}^{N} \left(i - \frac{N+1}{2} \right)^2 = \cdots = \frac{N(N+1)}{12}$$

  (4) Finally, compute the **Kruskal-Wallis Test Statistic**, denoted K, is defined to be

$$K := \frac{\text{SS}_\text{B}}{N(N+1)/12}$$

  Under the null, $K \sim \chi^2_{I-1}$. Thus, form a rejection region using the quantiles of the χ^2_{I-1} distribution, and complete the hypothesis test accordingly.

7 χ^2 Theory and Tests

- **χ^2 Goodness of Fit Test:** Could the data we observed plausibly have come from some distribution (let's call it $\mathcal{F}$)? To answer this, we use a χ^2 Goodness of Fit test. The null hypothesis is:

$$H_0 : \text{The } x_i\text{'s come from the distribution } \mathcal{F}$$

  - Let $\{x_i\}_{i=1}^m$ denote the sample values we observed. Let $\hat{x}_i$ denote the value of x_i that is predicted by the distribution $\mathcal{F}$. Then, the **Pearson χ^2 statistic** takes the form

$$X^2 := \sum_{i=1}^{m} \frac{(x_i - \hat{x}_i)^2}{\hat{x}_i}$$

  and, under the null with appropriately large sample size m, follows a χ^2 distribution with degrees of freedom equal to

$$\text{df} = m - 1 - \#\text{ parameters}$$

  Construct a rejection region according to the appropriate χ^2 distribution, and complete the hypothesis test.

  - As a quick example: could the following observations have come from a Poisson distribution?

$$1 \quad 1 \quad 0 \quad 0 \quad 2$$

  First arrange the data in the following manner:

# of Arrivals	0	1	2
Count	2	2	1

  Now, we fit a Poisson distribution to our data: we estimate the rate of this Poisson distirubtion using the MLE of λ, $\hat{\lambda}_{\text{ML}} = \text{sample mean}$. Therefore, our best guess for the Poisson distribution that fits this data is a Pois(0.8) distribution, and our null hypothesis is

$$H_0 : \text{The } x_i\text{'s come from a Pois(0.8) distribution}$$

  We can now compute expected counts. That is, letting $X \sim \text{Pois}(0.8)$, the expected number of k's (where $k \in \{0, 1, 2\}$) is

$$n \cdot \mathbb{P}(X = k) = 5 \cdot \cdot e^{-0.8} \frac{(0.8)^k}{k!}$$

  We thus fill in our table accordingly:

# of Arrivals	0	1	2
Observed	2	2	1
Expected	2.246645	1.797316	0.7189263

  and the Pearson χ^2 statistic is

$$X^2 = \frac{(2 - 2.246645)^2}{2.246645} + \frac{(2 - 1.797316)^2}{1.797316} + \frac{(1 - 0.7189263)^2}{0.7189263} \approx 0.1598$$

  Under H_0, X^2 will follow a χ^2 distribution with degrees of freedom $5 - 1 - 1 = 3$ (since we estimated one parameter, namely, λ): thus, the $p-$value of our test is 0.0162. Therefore, there is sufficient evidence to *reject* the null; we have reason to believe the data was *not* sampled from a Poisson distribution. [Of course, with such a small sample size we should be wary of any results; however, the main point of this was just to lay out the *procedure* for conducting a Goodness of Fit test.]

- **χ^2 Test of Homogeneity:** Given J different multinomial distributions, each with I observations, can we claim a statistical difference between these distributions? To answer this question, we use a χ^2 Test of Homogeneity. Letting π_{ij} denote the probability of the ith observation in the jth multinomial, we can phrase the null hypothesis as:

$$H_0 : \pi_{i1} = \pi_{i2} = \cdots = \pi_{iJ}, \quad i = 1, \ldots, I$$

  - In a sense, we will perform a Goodness of Fit Test, testing whether the distribution under the null is a good fit for our data. Specifically, the distribution under the null will be a Multinomial distribution where each of the J multinomials have the same probability π_i for the ith observation.
  - Let $n_{i\cdot}$ denote the number of responses in the ith category, and let $n_{\cdot\cdot}$ denote the total number of responses. Then, under the null, each of $\pi_1, \pi_2, \ldots, \pi_I$ has an MLE equal to

$$\hat{\pi}_i = \frac{n_{i\cdot}}{n_{\cdot\cdot}}, \quad i = 1, \ldots, I$$

    The expected count for the $(i,j)^{\text{th}}$ cell is

$$E_{ij} = n_{\cdot j} \times \hat{\pi}_i = \frac{n_{\cdot j}}{n_{i\cdot} n_{\cdot\cdot}}$$

    and, letting n_{ij} denote the $(i,j)^{\text{th}}$ observation, the Pearson χ^2 statistic takes the form

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{\cdot j}}{n_{i\cdot} n_{\cdot\cdot}}\right)^2}{\frac{n_{\cdot j}}{n_{i\cdot} n_{\cdot\cdot}}}$$

    When the sample size is very large, X^2 is approximately distributed as a χ^2 distribution with degrees of freedom

$$\text{df} = J(I-1) - (I-1) = (I-1)(J-1)$$

- **χ^2 Test of Independence:** Given J different multinomial distributions, each with I observations, can we conclude each of the J multinomial distributions to be independent? To answer this question, we use a χ^2 Test of Independence. Let π_{ij} denote the probability associated with the $(i,j)^{\text{th}}$ cell, and define the following marginal probabilities as follows:

$$\pi_{i\cdot} := \sum_{j=1}^{J} \pi_{ij} = \text{marginal probability that an observation will fall in the } i^{\text{th}}\text{row}$$

$$\pi_{\cdot j} := \sum_{i=1}^{I} \pi_{ij} = \text{marginal probability that an observation will fall in the } j^{\text{th}}\text{column}$$

  Our null hypothesis, namely that rows and columns are independent of each other, can then be phrased as

$$H_0 : \pi_{ij} = \pi_{i\cdot} \times \pi_{\cdot j}, \quad i = 1, \ldots, I;\ j = 1, \ldots, J$$

  - Under H_0, the MLE $\hat{\pi}_{ij}$ of π_{ij} is

$$\hat{\pi}_{ij} = \widehat{\pi_{i\cdot} \times \pi_{\cdot j}} = \hat{\pi}_{i\cdot} \times \hat{\pi}_{\cdot j} = \frac{n_{i\cdot}}{n} \times \frac{n_{\cdot j}}{n} = \frac{n_{i\cdot} n_{\cdot j}}{n^2}$$

    where n denotes the total number of observations, $n_{i\cdot}$ denotes the number of observations in row i, and $n_{\cdot j}$ denotes the number of observations in column j. The expected count for the $(i,j)^{\text{th}}$cell is

$$E_{ij} = n \times \hat{\pi}_{ij} = n \times \frac{n_{i\cdot} n_{\cdot j}}{n^2} = \frac{n_{i\cdot} n_{\cdot j}}{n}$$

and so, under H_0, the Pearson χ^2 statistic takes the form

$$X^2 = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{(\text{Observed}_{ij} - \text{Expected}_{ij})^2}{\text{Expected}_{ij}} = \sum_{i=1}^{I}\sum_{j=1}^{J} \frac{\left(n_{ij} - \frac{n_{\cdot j}}{n_{i\cdot}n}\right)^2}{\frac{n_{\cdot j}}{n_{i\cdot}n}}$$

When the sample size is very large, X^2 is approximately distributed as a χ^2 distribution with degrees of freedom

$$\text{df} = \underbrace{IJ}_{\#\text{ of observations}} - 1 - \underbrace{(I-1)-(J-1)}_{\#\text{ of parameters estimated}} = (I-1)(J-1)$$

- **What's the Difference between Tests of Homogeneity and Tests of Independence?** You'll note that the final *math* (as in, the test statistic and its distribution under the null) is the same for both Tests of Homogeneity and Tests of Independence. So, what's the difference between the two?

  - In a χ^2 test of homogeneity, we have two or more populations but only one categorial variable. In a χ^2 test of independence, however, we have only one population but two categorical variables. As such, the key difference between the two tests lies in the setup of the problem.

  - **Scenario 1:** I seek out 50 out-of state students and 100 in-state students at a University, note their favorite color and note whether they are an out-of-state student or not. Subsequently, I want to know if Favorite Color and Out-Of-State-ness are related.

    Here we have two populations (namely, Out-of-State and In-State) and only one categorical variable (namely, Favorite Color); as such, we should perform a test of homogeneity.

  - **Scenario 2:** I sample 150 students at a University, note their favorite color and note whether they are an out-of-state student or not; it turns out that 50 people in this sample are out-of-state, and the remaining 100 are in-state. Subsequently, I want to know if Favorite Color and Out-Of-State-ness are related.

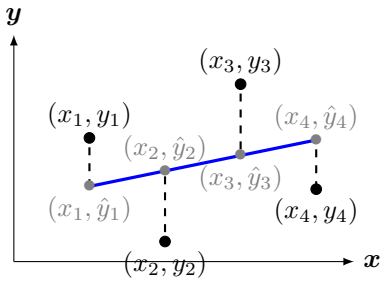
    Here we have only one population (namely 150 students) and two categorical variables (in/out-of-state-ness and Favorite Color); as such, we should perform a test of independence.

- **What's the Difference between χ^2 Tests and ANOVA?** On the surface, it seems as though both ANOVA and χ^2 tests are designed to answer the question "do relationships exist across these J groups I've observed?" **The key difference lies in the *type* of data being observed.**

  - In ANOVA, we assume continuous data (e.g. levels of drug in the bloodstream, weight measurements, etc.)
  - In χ^2 Tests of Homogeneity and Tests of Independence, we assume discrete data (counts of some sort). [Additionally, since the data is discrete, we *definitely* can't assume normally distributed data...]

8 Regression

- **Warm-Up: Line-Fitting using SSE** Consider the problem of fitting a line to a series of n points on the (x, y) plane. More specifically: given datapoints $\{(x_i, y_i)\}_{i=1}^n$, we associate with each point (x_i, y_i) another point $(x_i, \hat{y}_i)$ such that all the $\{(x_i, \hat{y}_i)\}_{i=1}^n$ lie along a line:



We need some way of assessing how "good" our line of fit is: ideally, we want our line to be as "close" to the datapoints as possible.

Define the **Sum of Squared Errors**, or SSE as follows: given a series of predictions $\{\hat{y}_i\}_{i=1}^n$ for values $\{y_i\}_{i=1}^n$, we define the SSE to be

$$\text{SS}_\text{E} := \sum_{i=1}^{n} (y_i - \hat{y}_i)^2$$

Graphically, this corresponds to the sum of squared-differences between the predicted values and the true values. If our fit is to be considered "good," we want SS_E to be as low as possible.

Mathematically, this means we want to *minimize* SS_E. However, what variable should we minimize over? Recall that our $\hat{y}_i$ values all fall along a line; as such, we may express our $\hat{y}_i$ as

$$\hat{y}_i = a + bx_i$$

Hence, we can actually minimize the SSE over a and b, the slope and intercept of our fitted line.

$$\begin{aligned}
\text{SS}_\text{E} &= \sum_{i=1}^{n} [y_i - (a + bx_i)]^2 \\
\frac{\partial}{\partial a}[\text{SS}_\text{E}] &= -2 \sum_{i=1}^{n} (y_i - a - bx_i) \stackrel{!}{=} 0 \\
\frac{\partial}{\partial b}[\text{SS}_\text{E}] &= -2 \sum_{i=1}^{n} x_i (y_i - a - bx_i) \stackrel{!}{=} 0
\end{aligned}$$

Solving the resulting system of equations for a and b yield the desired result.

- **Tying Together SSE Together and Regression.** Suppose, now, we have a set of n points $\{(x_i, y_i)\}_{i=1}^n$ where X and Y are related variables: for example, X could be "height" and Y could be "weight." We call Y the **response variable**, and we call X the **predictor variable**.
  - **Assumption:** We will assume that our response variable can be modeled as follows:

    $$y_i = \beta_0 + \beta_1 x_i + \varepsilon_i$$

    That is, we assume that our response variable and predictor variable are linearly related, plus or minus some random noise ε_i. Because this noise is random, its exact value at any instance i is unknown; the best we can try to do is fit the "de-noised" line $\beta_0 + \beta_1 x_i$ to our data.
  - Now that we have reduced our problem into that of line-fitting, we can use the results from the introductory warm-up above. Set

    $$S(\beta_0, \beta_1) := \text{SS}_\text{E} := \sum_{i=1}^{n} (y_i - \beta_0 - \beta_1 x_i)^2$$

and minimize $S(\beta_0, \beta_1)$ over β_0 and β_1, separately. Taking appropriate derivatives and setting equal to zero, we obtain the minimizing quantities

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{\sum_{i=1}^{n}(x_i - \bar{x})^2}; \qquad \hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{x}$$

Another useful formulation of $\hat{\beta}_1$ is

$$\hat{\beta}_1 = r \cdot \left(\frac{s_x}{s_y}\right)$$

where s_x is the sample variance of the predictor values, s_y is the sample variance of response variables, and r is the sample correlation between X and Y.

- **Multiple Regression:** Suppose instead of having only one predictor variable, we had p predictors, so that our data took the form $\{(x_{i1}, x_{i2}, \ldots, x_{ip}, y_i)\}_{i=1}^{n}$.
  - We will therefore use the model

$$y_i = \beta_0 + \beta_1 x_{i1} + \cdots + \beta_p x_{ip} + \varepsilon_i$$

This is more compactly expressed in matrix/vector form. Define the following quantities:

$$\vec{\boldsymbol{y}} = \begin{pmatrix} y_1 \\ \vdots \\ y_n \end{pmatrix}; \ \vec{\boldsymbol{\beta}} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_p \end{pmatrix}; \ \vec{\boldsymbol{\varepsilon}} = \begin{pmatrix} \varepsilon_1 \\ \vdots \\ \varepsilon_n \end{pmatrix}; \ \mathbf{X} = \begin{pmatrix} 1 & x_{11} & x_{12} & \ldots & x_{1p} \\ 1 & x_{21} & x_{22} & \ldots & x_{2p} \\ \vdots & \vdots & \ddots & \vdots & \\ 1 & x_{n1} & x_{n2} & \ldots & x_{np} \end{pmatrix}$$

In other words, the $(i,j)^{\text{th}}$element of $\mathbf{X}$ is the i^{th}observation of the j^{th}predictor variable. Then, our model can be neatly summarized as

$$\vec{\boldsymbol{y}} = \mathbf{X}\vec{\boldsymbol{\beta}} + \vec{\boldsymbol{\varepsilon}}$$

  - Our goal, therefore, is to fit a *hyperplane* $\mathbf{X}\hat{\boldsymbol{\beta}}$ to our data. The notion of SSE extends to multiple dimensions:

$$\text{SS}_\text{E} := \sum_{i=1}^{n}(y_i - \beta_0 - \beta_1 x_{i1} - \beta_2 x_{i2} - \cdots - \beta_p x_{ip})^2; \qquad i = 1, 2, \ldots, n$$

We could differentiate this $p+1$ times, to $p+1$ equations to estimate the $p+1$ unknowns $\beta_0, \beta_1, \ldots, \beta_p$; however, it will be much more efficient to write the SSE in vector notation as well:

$$S(\vec{\boldsymbol{\beta}}) := \text{SS}_\text{E} := \left\|\vec{\boldsymbol{y}} - \mathbf{X}\vec{\boldsymbol{\beta}}\right\|^2 = (\vec{\boldsymbol{y}} - \mathbf{X}\vec{\boldsymbol{\beta}})^\mathsf{T}(\vec{\boldsymbol{y}} - \mathbf{X}\vec{\boldsymbol{\beta}})$$

where $\|\cdot\|^2$ denotes the squared-Euclidean norm. Expanding our expression for $S(\vec{\boldsymbol{\beta}})$, differentiating (with respect to a vector!!!) and setting equal to $\vec{\mathbf{0}}$ yields the following equation which $\hat{\boldsymbol{\beta}}$ must satisfy:

$$(\mathbf{X}^\mathsf{T}\mathbf{X})\hat{\boldsymbol{\beta}} = \mathbf{X}^\mathsf{T}\vec{\boldsymbol{y}} \qquad \text{(Normal Eqns.)}$$

These are known as the **Normal Equations**. (Note that we use equation*s*, plural, because this is actually a system of $p+1$ equations in $p+1$ unknowns!) Assuming $\mathbf{X}^\mathsf{T}\mathbf{X}$ is invertible, we find

$$\hat{\boldsymbol{\beta}} = (\mathbf{X}^\mathsf{T}\mathbf{X})^{-1}\mathbf{X}^\mathsf{T}\vec{\boldsymbol{y}} \qquad \text{(OLS Estimate)}$$

which is known as the **Ordinary Least Squares Estimate of $\vec{\boldsymbol{\beta}}$** (or OLS estimate, for short).

- **Prediction:** Suppose that we now obtain a *new* set of observations on the p predictor variables: $(x_{n+1,1}, x_{n+1,2}, \ldots, x_{n+1,p})$. Set

$$\vec{\boldsymbol{x}}_{n+1} = \begin{pmatrix} 1 \\ x_{n+1,1} \\ x_{n+1,2} \\ \vdots \\ x_{n+1,p} \end{pmatrix}$$

  Then, our best estimate for the true response variable y_{n+1} associated with these new observations is

$$\hat{y}_{n+1} = \vec{\boldsymbol{x}}_{n+1}^{\mathsf{T}} \hat{\boldsymbol{\beta}}$$

- We may construct a $(1-\alpha) \times 100\%$ confidence interval for the mean $Y-$value given an an observation $\vec{\boldsymbol{x}}_0$ using

$$\vec{\boldsymbol{x_0}}^{\mathsf{T}} \hat{\boldsymbol{\beta}} \pm t_{n-p-1}^{(\alpha/2)} \cdot \hat{\sigma} \cdot \sqrt{\vec{\boldsymbol{x_0}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \vec{\boldsymbol{x_0}}}$$

  If instead we seek to construct an interval for an *individual* observation of y, we instead construct a **prediction interval** using

$$\vec{\boldsymbol{x_0}}^{\mathsf{T}} \hat{\boldsymbol{\beta}} \pm t_{n-p-1}^{(\alpha/2)} \cdot \hat{\sigma} \cdot \sqrt{1 + \vec{\boldsymbol{x_0}}^{\mathsf{T}} (\mathbf{X}^{\mathsf{T}}\mathbf{X})^{-1} \vec{\boldsymbol{x_0}}}$$

  where p denotes the number of explanatory variables, and $t_{n-p-1}^{(\alpha/2)}$ denotes the $\alpha/2^{\text{th}}$quantile of the t_{n-p-1} distribution. Typically, to estimate σ we use

$$\hat{\sigma} = \frac{\text{RSS}}{n-(p+1)}$$

- **Should I Include an Intercept or Not?** Note that the general model above assumes the presence of an intercept β_0. In certain cases, however, it makes sense to *omit* the intercept; usually this will be obvious given the context of a problem. That is, if we have an observed $x-$value of 0, it will typically be obvious whether our response value should be 0 (in which case we *omit* the intercept from our model) or not (in which case we *include* the intercept from our model).

9 Bayesian Statistics

- In the **Bayesian** framework, population parameters are considered to be *random variables*, rather than fixed deterministic quantities (as is the case in the **frequentist** framework).
  - This enables us to **update the prior**; intuitively, this refers to the idea of updating our viewpoint of the world every time new information is given to us.
- An illustrative example may be useful: consider a two-sided coin that lands either heads or tails. In a Bayesian framework, we assume that the probability of the coin landing heads is a random variable Θ. The distribution of Θ is called the **prior distribution**, and is set before conducting any experiments.
  - Picking a prior distribution isn't always easy, though often times there are physical constraints that aid us in our choice of a prior. For example, in the coin-tossing scenario, Θ must have a support of $[0,1]$. As such, a natural choice of prior might be to set $\Theta \sim \text{Unif}[0,1]$, or $\Theta \sim \text{Beta}(\alpha, \beta)$ where α and β are real numbers (called **hyperparameters**).
  - Sometimes, priors are set in a certain way because they simplify math; see the notes on *conjugacy* below.

- At the heart of Bayesian Statistics is a form of **Bayes' Rule**, which states

$$f_{\Theta|X}(\theta \mid x) = \frac{f_{X|\Theta}(x \mid \theta) \times f_\Theta(\theta)}{f_X(x)} = \frac{f_{X|\Theta}(x \mid \theta) \times f_\Theta(\theta)}{\int_\Theta f_{X|\Theta}(x \mid \tilde{\theta}) \times f_\Theta(\tilde{\theta}) \, \mathrm{d}\tilde{\theta}}$$

  A key to Bayesian Statistics is becoming comfortable with **ignoring normalizing constants**. For example, in the ratio above, the **marginal** $f_X(x)$ doesn't give us any new information. As such, we generally consider the following representation:

$$\underbrace{f_{\Theta|X}(\theta \mid x)}_{\textbf{Posterior}} \propto \underbrace{f_{X|\Theta}(x \mid \theta)}_{\textbf{Likelihood}} \times \underbrace{f_\Theta(\theta)}_{\textbf{Prior}}$$

- A prior and a posterior are said to be in **conjugacy** (or, stated differently, they are said to form a **conjugate pair**) if both the prior and posterior belong to the same family of distributions.
  - **Beta-Binomial Conjugacy:** If we choose a beta prior and a binomial likelihood, then the posterior will also follow a beta distribution.
  - If you have an idea of what the posterior distribution might be, it may make sense to set the prior to be a conjugate pair with the posterior (so that the algebra simplifies considerably).

- The **Maximum A Posteriori Estimator** (or **MAP** estimator) of a parameter Θ can be loosely thought of as the Bayesian analog of MLE:

$$\hat{\theta}_{\text{MAP}} := \underset{\theta}{\text{argmax}} \left\{ f_{\Theta|X}(\theta \mid x) \right\}$$

  In other words, it is the maximum value (i.e. mode) of the posterior distribution, and represents a notion of the "most probable" value of Θ.

- It will be instructive to complete an example. Set $\Theta \sim \text{Unif}[0,1]$ and $(X \mid \Theta = \theta) \sim \text{Bin}(n, \theta)$. Then, in this example (assuming a sample $\{x\}$ of size 1),

$$\text{Prior Density: } f_\Theta(\theta) = \mathbb{1}\{\theta \in [0,1]\}$$

$$\text{Likelihood Density: } \mathbb{P}(X = x \mid \Theta = \theta) = \binom{n}{x} \theta^x (1-\theta)^{n-x}$$

  Then, the posterior density is obtained by

$$f_{\Theta|X}(\theta \mid x) \propto \binom{n}{x} \theta^x (1-\theta)^{n-x} \cdot \mathbb{1}\{\theta \in [0,1]\} \propto \theta^x (1-\theta)^{n-x} \cdot \mathbb{1}\{\theta \in [0,1]\}$$

  Notice that we dropped the initial binomial factor; this is because it doesn't give us any information about θ, and will be absorbed by the normalizing constant appearing in the denominator of Bayes' Rule. This shows that

$$\boxed{(\Theta \mid X = x) \sim \text{Beta}(x+1, n+1-x)}$$

  To find the MAP estimate of Θ we differentiate the likelihood, with respect to θ:

$$\frac{\mathrm{d}}{\mathrm{d}\theta} f_{\Theta|X}(\theta \mid x) \propto x\theta^{x-1}(1-\theta)^{n-x} - \theta^x (n-x)(1-\theta)^{n-x-1}$$

  Set equal to 0 to see

$$x\hat{\theta}^{x-1}(1-\hat{\theta})^{n-x} = \hat{\theta}^x (n-x)(1-\hat{\theta})^{n-x-1}$$

  or, equivalently,

$$\frac{x}{\hat{\theta}} = \frac{n-x}{1-\hat{\theta}}$$

  which implies that $\boxed{\hat{\theta}_{\text{MAP}} = x/n}$.

10 Miscellaneous Concepts

Note: These are concepts that are not explicitly taught in Stat 135, however they are crucial for *all* 150 series classes. As such, we believe it will be beneficial for you to read this section.

10.1 Convergence: A Deeper Dive

- First recall the two types of convergence, and their definitions:
  - A sequence $\{X_i\}_{i=1}^{\infty}$ converges in probability to a random variable X if, for every $\varepsilon > 0$, we have

$$\lim_{n\to\infty} \mathbb{P}\left(|X_n - X| < \varepsilon\right) = -$$

    In other words, convergence in probability means that, as n gets very large, we become more and more certain that X_n and X are really the same thing. We typically notate convergence in probability by:
  - A sequence $\{X_i\}_{i=1}^{\infty}$ **converges in distribution** to a random variable X if

$$\lim_{n\to\infty} F_{X_n}(x) = F_X(x)$$

    where $F_{X_n}(\cdot)$ denotes the c.d.f. of X_n, and $F_X(\cdot)$ denotes the c.d.f. of X. We typically notate convergence in probability by:

$$X_n \xrightarrow{\text{d}} X$$

- Convergence in probability is a **stronger** condition than convergence in distribution; in fact,

$$(X_n \xrightarrow{\text{p}} X) \implies (X_n \xrightarrow{\text{d}} X)$$

  The converse is not necessarily true.

- **Central Limit Theorem, Revisited:** The CLT is actually a statement about convergence in distribution! Here is a more formal phrasing of its statement: Let $\{X_i\}_{i=1}^{\infty}$ denote a sequence of independent random variables with mean μ and variance σ^2. Define

$$S_n := \sum_{i=1}^{n} X_i$$

  Then,

$$\left(\frac{S_n - n\mu}{\sigma\sqrt{n}}\right) \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

  Phrased slightly differently,

$$\lim_{n\to\infty} \mathbb{P}\left(\frac{S_n - n\mu}{\sigma\sqrt{n}} \le x\right) = \Phi(x) := \int_{-\infty}^{x} \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}t^2}\, \mathrm{d}t$$

- Recall, from Stat 134, we say that the normal distribution can "approximate" the binomial distribution. In fact, we can phrase this more rigorously using convergence concepts: If $X_n \sim \text{Bin}(n,p)$ and $\hat{p}_n := X_n/n$, we have

$$\frac{X_n - np}{\sqrt{np(1-p)}} \xrightarrow{\text{d}} \mathcal{N}(0,1); \qquad \frac{\hat{p}_n - p}{\sqrt{p(1-p)/n}} \xrightarrow{\text{d}} \mathcal{N}(0,1)$$

10.2 Establishing Consistency

When trying to prove that an estimator is consistent, the following theorem may be useful:

Establishing Consistency: Let $\hat{\theta}_n$ be an unbiased estimator of θ. If $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then $\hat{\theta}_n$ is a consistent estimator of θ.

Proof: Since $\hat{\theta}_n$ is an unbiased estimator of θ, we have that $\mathbb{E}(\theta_n) = \theta$. Let $\sigma_{\hat{\theta}_n}$ denote the standard deviation of $\hat{\theta}_n$ (for a fixed sample size n); then, for $k > 0$, Chebyshev's Inequality tells us that

$$0 \leq \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq k\sigma_{\hat{\theta}_n}\right) \leq \frac{1}{k^2} \tag{1}$$

Fix an arbitrary $\varepsilon > 0$; since equation 1 holds for any $k > 0$, it must hold for $k = \varepsilon/\hat{\theta}_n$, which yields

$$0 \leq \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \frac{\varepsilon}{\cancel{\sigma_{\hat{\theta}_n}}} \cdot \cancel{\sigma_{\hat{\theta}_n}}\right) \leq \frac{\sigma^2_{\hat{\theta}_n}}{\varepsilon^2}$$

Recognizing $\sigma^2_{\hat{\theta}_n} = \text{Var}(\hat{\theta}_n)$, and simplifying terms, we obtain

$$0 \leq \mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \leq \frac{\text{Var}(\hat{\theta}_n)}{\varepsilon^2}$$

Now, take the limit as $n \to \infty$; if $\text{Var}(\hat{\theta}_n) \to 0$ as $n \to \infty$, then the squeeze theorem (from mathematics) tells us that

$$\mathbb{P}\left(\left|\hat{\theta}_n - \theta\right| \geq \varepsilon\right) \to 0$$

which is precisely the definition of the statement that $\hat{\theta}_n$ is a consistent estimator of θ. □

11 Additional Resources

All of Statistics, by Larry J. Wasserman (Available for free through SpringerLink)

Though concise, this book provides a good overview of many concepts of Probability and Statistics. Specifically, chapters 5, 6, 8, 9, 10, and 11 all have sections that are directly applicable to Stat 135. We highly recommend this text as a refresher of concepts, as well as a resource for future coursework.

Mathematical Statistics with Applications, by Wackerly, Mendenhall, and Scheafer

This text is quite similar to Rice in may ways, but with a slightly different organization of topics. Additionally, this text contains a more rigorous discussion of convergence topics as well as a more rigorous introduction to Bayesian Statistics. Several universities have adopted this as a textbook of choice; chapters 7 - 16 contain material directly applicable to Stat 135.

For those with a slightly higher mathematical curiosity:
Statistical inference, by Casella and Berger

This is a popular text among graduate-level mathematical statistics classes, though many sections are quite well-written and comprehensible by a motivated undergraduate student. If ever you find yourself curious on how to dive deeper into a particular subject, we recommend referencing this textbook.