15. Statistics

Introduction

Probability. What kinds of questions are addressed in probability?

- The questions so far in the course about cards, dice, birthdays etc.
- For example, what is the chance that two people in this class were born on the same day?

Statistics. What kinds of questions are addressed in statistics? Here are some examples.

- We want to know the proportion of left-handed and right-handed people in Lausanne. How many people need to be surveyed to achieve an estimate with $\pm x\%$ accuracy, valid at 99% confidence?
- We have data from a large experiment, run by Novartis, where we have one group who received a cancer drug and one group who received control. How do we use the results of the experiment to conclude that the experiment has an effect?

Probability vs. Statistics? To simplify:

- **In probability**, we know the model and focus on what it generates.
- **In statistics**, we have data and focus on the underlying model (in reality, it's more complex). In that sense, statistics is the science of learning from data.

Thus:

- The probabilist poses their model F, and uses probability laws to *deduce* the properties of Y—they are certain, if their reasoning is correct!
- The statistician does the opposite: they use the data y to *infer* the properties of the model F—they are uncertain, because y is finite and it is rarely certain that their assumptions are correct.

- Key points:

  - The *context* of the problem is important—it is essential to know how the data were collected and what they represent when doing statistics; this determines the assumptions we make when doing statistical modelling
  - The *variability* (of the data) and the *uncertainty* that result are represented with probabilistic models;
  - We try to *quantify uncertainty* when drawing conclusions (we will see how to do this in terms of confidence intervals), and
  - *account for uncertainty* when choosing actions based on a study.

Probabilities and Data

- To connect data and probabilities, suppose our observations $y = (y_1, y_2, \ldots, y_n)$ are random:
  - either by imposing a *random mechanism*, such as the *randomization* of an experiment or a survey;
  - or by assuming they result from a *random process*, e.g., suppose that the delay R of my bus follows an $\exp(\lambda)$ law, and I try to estimate $P(R > 5)$ from observations $r_1, \ldots, r_n$, because I want to arrive on time for a lecture...
- Often, we study the behavior of a variable y in
  - a *population*—the entire set of interest for our investigation—from which we sample
  - a *sample* $y_1, \ldots, y_n$,
  - assuming this sample is a realization of random variables $Y_1, \ldots, Y_n$ from a *probabilistic model* F.

Statistical Process

Main steps are:

- *Formulating* the research question or hypotheses;
- *Data collection*, leading to
  - *Experiment planning* (design, implementation, and data acquisition);
  - If experiments are not possible, an *observational study*, where the data collection is not under the investigator's control.
- *Data analysis*, Two different ways:
  - Exploratory analysis,
  - **Inference.** (which is what I will focus on).
- *Interpreting results and drawing practical conclusions.*

Nomenclature

- When we use the word law, we are referring to a distribution, which is characterized by the CDF (or, equivalently with the PDF or PMF).
- **Definition 134.** A *statistical model* is a law (or family of laws) of probability constructed for a statistical study. A *parameter* is any (constant) function of a CDF, often denoted with Greek letters. A model determined by a finite-dimensional parameter is *parametric*, otherwise, it is *nonparametric*.
- **Definition 135.** A *statistic* $S = s(Y)$ is any function of data Y. This includes functions like the mean, but also graphs.
- **Definition 136.** The *sampling distribution* of a statistic $S = s(Y)$ is its probability law when Y is generated by a statistical model.
- **Definition 137.** A *random sample* $Y_1, \ldots, Y_n \overset{\text{iid}}{\sim} F$ is a realization $y_1, \ldots, y_n$ of such $Y_1, \ldots, Y_n$.

Concept	Description
Estimand	The quantity or parameter of interest that we aim to learn about from the data, defined as a function of the population's probability distribution or CDF. Example: the population mean $\mu = E[Y]$.
Estimator	A rule, algorithm, or function that maps the observed data Y to an estimate of the estimand. It is a statistic $S = s(Y)$ derived from the data. Example: $\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} Y_i$.
Estimate	The numerical value obtained from the estimator when applied to observed data. Example: $\hat{\mu} = 5.2$ based on a specific dataset.

TABLE 4. Difference between Estimand, Estimator, and Estimate

Example: Evaluating effects of a drug

Drug company test a new drug on elderly people with breast cancer. They only have a **sample** of n individuals, where n_1 is the number of treated and n_0 is the number of untreated individuals. The aim is to assess the effect of the drug on being cured.

We use "hats" to denote estimates below.

- **Step 1: Compute Proportions in Each Group**
  - Treated group ($\hat{p}_1$): The proportion cured in the treated group is:

$$\hat{p}_1 = \frac{\text{Number of Cured in Treated Group}}{\text{Total Number in Treated Group}} = \frac{300}{500} = 0.6.$$

  - Untreated group ($\hat{p}_0$): The proportion cured in the untreated group is:

$$\hat{p}_0 = \frac{\text{Number of Cured in Untreated Group}}{\text{Total Number in Untreated Group}} = \frac{250}{500} = 0.5.$$

    Q: What does the LLN tell us about these estimates?

    Q: What does the central limit theorem say about the distribution of these estimates?

- **Step 2: Compute Confidence Intervals for Each Proportion**
  - Treated group ($\hat{p}_1$): The 95% confidence interval is:

$$\text{CI} = \hat{p}_1 \pm z \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1}},$$

    where $z = 1.96$, $\hat{p}_1 = 0.6$, and $n_1 = 500$. Substituting:

$$\text{CI} = 0.6 \pm 1.96 \cdot \sqrt{\frac{0.6 \cdot 0.4}{500}} = 0.6 \pm 1.96 \cdot 0.022.$$

$$\text{CI} = (0.556, 0.644).$$

  - Untreated group ($\hat{p}_0$): The 95% confidence interval is:

$$\text{CI} = \hat{p}_0 \pm z \cdot \sqrt{\frac{\hat{p}_0(1-\hat{p}_0)}{n_0}},$$

where $z = 1.96$, $\hat{p}_0 = 0.5$, and $n_0 = 500$. Substituting:

$$\text{CI} = 0.5 \pm 1.96 \cdot \sqrt{\frac{0.5 \cdot 0.5}{500}} = 0.5 \pm 1.96 \cdot 0.022.$$

$$\text{CI} = (0.456, 0.544).$$

- **Step 3: Compute the Difference in Proportions**
  - The difference in proportions $(\hat{p}_1 - \hat{p}_0)$ is:

$$\hat{p}_1 - \hat{p}_0 = 0.6 - 0.5 = 0.1.$$

- **Step 4: Confidence Interval for the Difference in Proportions**
  - Formula: The 95% confidence interval for the difference in proportions is:

$$\text{CI} = (\hat{p}_1 - \hat{p}_0) \pm z \cdot \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_0(1-\hat{p}_0)}{n_0}}.$$

  - Substituting values:

$$\text{CI} = 0.1 \pm 1.96 \cdot \sqrt{\frac{0.6 \cdot 0.4}{500} + \frac{0.5 \cdot 0.5}{500}} = 0.1 \pm 1.96 \cdot 0.031.$$

$$\text{CI} = (0.1 - 0.061, 0.1 + 0.061) = (0.039, 0.161).$$

- **Step 5: Relation to the CLT and LLN**
  - **Central Limit Theorem (CLT):** The CLT ensures that the sample proportion $(\hat{p})$ is approximately normally distributed for large sample sizes (n). This allows us to use the normal approximation to construct confidence intervals for the sample proportion $(\hat{p})$.
  - **Law of Large Numbers (LLN):** The LLN guarantees that the sample proportion $(\hat{p})$ converges to the true population proportion (p) as the sample size (n) increases. This justifies the reliability of the point estimates $(\hat{p})$.
- **Step 6: Final Results**
  - Confidence interval for the treated group $(\hat{p}_1)$: $(0.556, 0.644)$.

- Confidence interval for the untreated group ($\hat{p}_0$): $(0.456, 0.544)$.
- Confidence interval for the difference in proportions ($\hat{p}_1 - \hat{p}_0$): $(0.039, 0.161)$.

Interpretation of a Confidence Interval (CI)

- A fundamental goal of statistics is to provide an understanding of the uncertainty of an observation, and the confidence interval is one way of quantifying such uncertainty.
- Let θ be an unknown parameter, and let $\hat{\theta}$ be an estimate of θ based on data $y_1, \ldots, y_n$:
  - If $n = 10^5$, we can be much more confident that $\hat{\theta} \approx \theta$ than if $n = 10$.
  - To express this, we want to construct an interval that is wider when $n = 10$ and narrower when $n = 10^5$, clearly illustrating the uncertainty associated with $\hat{\theta}$.
- **Definition 163.** Let $Y = Y_1, \ldots, Y_n$ be data sampled from a distribution F with a scalar parameter θ of interest. A **confidence interval (CI)** (L, U) for θ is a statistic that takes the form of an interval and contains θ with a specified probability. This probability is called the **confidence level** of the interval.
- **Notes:**
  - The limits L and U are functions of the data $Y_1, \ldots, Y_n$, not unknown quantities.
  - A two-sided (bilateral) confidence interval, of the form (L, U), is most commonly used.
  - A one-sided (unilateral) confidence interval, of the form $(-\infty, U)$ or (L, ∞), can sometimes be useful.

Approximate Confidence Intervals

- In most cases, approximate confidence intervals (CIs) are constructed based on estimators where variance estimates are required; that is, we usually don't know the true variance.

Definition 164. Let $\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_n)$ be an estimator of θ, $\tau_n^2 = \text{var}(\hat{\theta})$ its variance, and $V = v(Y_1, \ldots, Y_n)$ an estimator of τ_n^2. The quantity $V^{1/2}$ (or its realization $v^{1/2}$) is called the standard error of $\hat{\theta}$.

- (L, U) is a random interval that contains the parameter θ with a specified probability, $1 - \alpha$.
- Imagine an infinite series of repetitions of the experiment, resulting in different (L, U) intervals.
- The CI we calculate is one of the possible CIs, and we can consider it as being randomly chosen among them.
- We do not know if our random CI (L, U) contains θ, but this event has a probability of $1 - \alpha$.

Hypothesis testing

- **Objective:** Evaluate whether observed data provides sufficient evidence to reject a null hypothesis (H_0) in favor of an alternative hypothesis (H_a).
- **Key Steps:**
  - **Define Hypotheses:**
    - Null Hypothesis (H_0): Assumes no effect or difference.
    - Alternative Hypothesis (H_a): Assumes an effect or difference exists.
  - **Test Statistic:** Calculate a statistic (e.g., t-statistic, z-statistic) that summarizes the data.
  - **Sampling Distribution:** Assume H_0 is true to derive the distribution of the test statistic.
  - **P-value:** Compute the probability of observing data as extreme as, or more extreme than, the observed data, under H_0.
  - **Decision Rule:** Reject H_0 if the p-value is less than a pre-specified significance level (α, often 0.05 but the choice is somewhat arbitrary).
- **Interpretation:**
  - If H_0 is rejected, the result is indicating evidence against H_0.
  - Failing to reject H_0 does not confirm it is true; it indicates insufficient evidence against it.
- **Example** of a null hypothesis to test:

$$H_0 : \text{The drug has no effect on the cure rate.}$$

  This approach involves attempting to refute the null hypothesis, which assumes no difference in cure rates between the treated and untreated groups—a "stochastic proof by contradiction."
- Data obtained:
  - Treated group: $n_1 = 500$, $\hat{p}_1 = 0.6$ (proportion cured).

  - Untreated group: $n_2 = 500$, $\hat{p}_2 = 0.5$ (proportion cured).

- Test statistic:

$$z_{test} = \frac{\hat{p}_1 - \hat{p}_2}{SE},$$

  where:

$$SE = \sqrt{\hat{p}(1-\hat{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)},$$

  and the pooled proportion is:

$$\hat{p} = \frac{x_1 + x_2}{n_1 + n_2} = \frac{300 + 250}{500 + 500} = 0.55.$$

  Substituting:

$$SE = \sqrt{0.55 \cdot 0.45 \cdot \left(\frac{1}{500} + \frac{1}{500}\right)} = \sqrt{0.55 \cdot 0.45 \cdot \frac{2}{500}} \approx 0.029.$$

  The test statistic becomes:

$$z_{test} = \frac{0.6 - 0.5}{0.029} \approx 3.45.$$

- Compute the p-value:

$$p_{\text{obs}} = 2 \cdot P(Z_{test} > |z_{test}|) = 2 \cdot P(Z_{test} > 3.45).$$

  Using the standard normal distribution:

$$p_{\text{obs}} \approx 2 \cdot 0.00028 = 0.00056.$$

- Interpret results:
  - Either H_0 is true, and the observed difference in cure rates is due to random chance, or:
  - H_0 is false, and the drug has a statistically significant effect on the cure rate.
- Decision:

  - Since $p_{\text{obs}} = 0.00056$ is much smaller than the significance level ($\alpha = 0.05$), we reject H_0.
  - This suggests strong evidence that the drug increases the cure rate.

- Additional note:
  - Had $p_{\text{obs}} \approx 0.05$, the result would have been less convincing, and further studies might have been necessary before concluding the drug's effectiveness.

16. Some claims about hypothesis testing and p-values

- The probability that the 95% CI from our study includes the true parameter is 95%. False
- If we repeat our study in many random samples from the same population, the 95% CI will include the true parameter in 95% of the samples.
- If we published hundred of studies, at the end of our career we expect that, the 95% CI included the true parameter in 95% of our studies
- The p-value for the null hypothesis is the probability that the test hypothesis is true. False
- The p-value for the null hypothesis is the probability that chance alone produced the observed association False, unless better specified. The p-value is a probability computed under the null hypothesis...
- The p-value for the null hypothesis is the probability of obtaining an estimate at least as far from the null as the estimate we have obtained, only if the null hypothesis and all other assumptions used to compute the p-value are true.

When Doing Inference, What Makes a Good Estimator?

- We now assume $Y \equiv Y_1, \ldots, Y_n \overset{iid}{\sim} F$ and that we want to estimate $\theta = \theta(F)$ with an estimator $\hat{\theta} = \hat{\theta}(Y_1, \ldots, Y_n)$.
- Two criteria for good estimators:
  - *Asymptotic*: How does $\hat{\theta}$ behave as $n \to \infty$?
  - *Finite-sample*: How to compare $\hat{\theta}_1$ and $\hat{\theta}_2$ for a fixed n?
- Consistency is a key asymptotic criterion: $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$.

Definition 152. An estimator $\hat{\theta}$ of θ is called *consistent* if $\hat{\theta} \xrightarrow{P} \theta$ as $n \to \infty$.

Mean Squared Error.

Definition 154. The *bias* $b(\hat{\theta}; \theta)$ of $\hat{\theta}$ is $b(\hat{\theta}; \theta) = E(\hat{\theta}) - \theta$.

Definition 155. The *mean squared error (MSE)* of $\hat{\theta}$ is

$$\mathrm{MSE}(\hat{\theta}; \theta) = E\left\{(\hat{\theta} - \theta)^2\right\}.$$

Lemma 156. We can write $\mathrm{MSE}(\hat{\theta}; \theta) = b(\hat{\theta}; \theta)^2 + \mathrm{var}(\hat{\theta})$.

- Bias is a property of the estimator $\hat{\theta}$, and this property may vary depending on θ.
- Interpretation of bias:
  - If $b(\hat{\theta}; \theta) < 0$, then on average $\hat{\theta}$ underestimates θ.
  - If $b(\hat{\theta}; \theta) > 0$, then on average $\hat{\theta}$ overestimates θ.
  - If $b(\hat{\theta}; \theta) = 0$, then $\hat{\theta}$ is *unbiased*.
- A quality indicator of $\hat{\theta}$ is the absence of systematic deviation from θ: $b(\hat{\theta}; \theta) \approx 0$.
- An even more important indicator is $\mathrm{MSE}(\hat{\theta}; \theta)$, which also measures the variability of $\hat{\theta}$.

Note to Example 157.

- We've already seen that

$$E(\overline{X}_n) = \frac{1}{n}E(X_1 + \cdots + X_n) = \frac{1}{n}(E(X_1) + \cdots + E(X_n)) = \mu,$$

so this mean estimator is unbiased.

Also, the sample variance

$$S_n^2 = \frac{1}{n-1}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$$

is unbiased because

$$\mathbb{E}(S_n^2) = \sigma^2.$$

.

However, had we rather made an alternative variance estimator

$$R_n^2 = \frac{1}{n}\sum_{j=1}^{n}(X_j - \overline{X}_n)^2.$$

we would have bias downwards because $E(R_n) < E(S_n)$.

Efficiency.

Definition 158. Let $\tilde{\theta}_1$ and $\tilde{\theta}_2$ be two unbiased estimators of the same parameter θ. Then

$$\text{MSE}(\tilde{\theta}_1; \theta) = \text{var}(\tilde{\theta}_1), \quad \text{MSE}(\tilde{\theta}_2; \theta) = \text{var}(\tilde{\theta}_2),$$

and we say that $\tilde{\theta}_1$ is more *efficient* than $\tilde{\theta}_2$ if

$$\text{var}(\tilde{\theta}_1) \leq \text{var}(\tilde{\theta}_2), \quad \theta \in \Theta.$$

We prefer $\tilde{\theta}_1$.

Example 159. Let $Y_1, \ldots, Y_n \overset{iid}{\sim} \mathcal{N}(\mu, \sigma^2)$, with large n. Find the properties of the sample median M_n and the sample mean $\bar{Y}_n$. Which is preferable? And what if outliers are present?

Example of unbiasedness.

- We've already seen that

$$E(\bar{Y}) = \mu, \quad \text{var}(\bar{Y}) = \sigma^2/n,$$

so the bias of $\bar{Y}$ as an estimator of μ is $E(\bar{Y}) - \mu = 0$.

- One can also show that for large n,

$$E(M) \approx \mu, \quad \text{var}(M) \approx \frac{\pi\sigma^2}{2n},$$

so both estimators are (approximately) unbiased (in fact exactly unbiased), but

$$\frac{\text{var}(M)}{\text{var}(\bar{Y})} = \frac{\pi}{2} > 1,$$

which means M is less efficient than $\bar{Y}$; its variance is larger. However, if there are outliers, the median M is little affected, while the mean $\bar{Y}$ can be significantly impacted. Our choice between these estimators will depend on how much we fear that our data will be contaminated by outliers.