Math 1620: Mathematical Statistics *Lecture Notes*

Y. Jain

Spring 2022

These are lecture notes for Math 1620: Mathematical Statistics taught at Brown University by Yajit Jain in the Spring of 2023.

These notes are taken by Jiahua Chen with gracious help and input from classmates. Please direct any mistakes/errata to me via email, or feel free to pull request the notes repository (https://github.com/jchen/math1620-notes).

Notes last updated February 23, 2023.

Contents

References

[Ber85] J.O. Berger. *Statistical Decision Theory and Bayesian Analysis*. Springer Series in Statistics. Springer, 1985.

[CB02] G. Casella and R.L. Berger. *Statistical Inference*. Duxbury advanced series in statistics and decision sciences. Thomson Learning, 2002.

[DS12] M.H. DeGroot and M.J. Schervish. *Probability and Statistics*. Addison-Wesley, 2012.

[Was04] L. Wasserman. *All of Statistics: A Concise Course in Statistical Inference*. Springer Texts in Statistics. Springer, 2004.

§1 January 26, 2022

§1.1 Introductions

We'll introduce this course and get an overview of the course.

Math 1610 is not a hard prerequisite, and there are enough people in the course that makes it flexible.

There will be a programming component of the course, which will either be in R or Python.

We'll jump right in!

§1.2 Probability

Definition 1.1 (Mathematical setting of probability and its meaning)
We have a probability space $(\Omega, \mathcal{A}, P)$ consisting of:

- a set Ω called the *sample space*,
- a σ-algebra $\mathcal{A}$ consisting of a selection of subsets of Ω,
- and a probability measure P.

Elements of Ω are called *outcomes*, subsets of Ω are called *events*, so that $\mathcal{A}$ is a collection of certain events, and P is a function

$$P : \mathcal{A} \to [0, 1]$$

that assigns to every event in $\mathcal{A}$ a number between 0 and 1.

In this context, we have

Definition 1.2 (Random Variables)

We have random variables, which are measurable functions

$$X : \Omega \to \mathbb{R}$$

Associated to a random variable X is the cumulative distribution function (cdf)

$$F_X(x) = P(X \leq x) = P(\{\omega \in \Omega : X(\omega) \leq x\}).$$

We distinguish between discrete and continuous random variables:

- Discrete random variables have a probability mass function (pmf)
- Continuous random variables have a probability density function (pdf)

This is the mathematical framework of probability theory. Every random variable is encoded in its pmf or pdf.

People coming from a pure math might want an *axiomatic framework* and to run with that. When you try to ask questions about events in the real world, you need an *extra* something. You want an extra computation (like a statistic) of whether an event is going to play out.

We'll be relying on a theoretical foundation and extract out numerics about what experiments in the real world are going to do based on some observations.

§1.3 Frequentist or Bayesian?

There are two main perspectives on the meaning of probability: frequentist and Bayesian.

Definition 1.3 (Frequentist Interpretation)

In the frequentist interpretation, the probability of an event is equal to the long-term frequency of the event's occurrence when the same process is *repeated* many times.

If we have a coin, we have $\frac{1}{2}$ probability of each heads or tails. When we flip that coin many times, we get roughly half heads and half tails.

Notice we had an *exact* number. In frequentist point of views, we do not attach probabilities to any hypotheses or fixed but unknown quantities, like p, θ, or λ in the definition of the standard distribution functions.

Example 1.4

Consider the statement "the probability of pulling an ace out of a deck of cards is $\frac{4}{52}$."

What is the frequentist explanation? If we used the same deck of cards, and took a card out of each of them (replacing and shuffling). Roughly $\frac{4}{52}$ of those trials (lots, like $10,000$) will result in an Ace.

What are problems in the real world? We might not get to do $10,000$ trials, like the weather.

Definition 1.5 (Bayesian Interpretation)

In the Bayesian interpretation, probabilities measure *degrees of belief*. These can be beliefs about the occurrence of an event, the truth of a hypothesis, or the truth of any random fact.

We're trying to measure the uncertainty of a hypothesis. With this definition, you can talk about the probabilities of single events, which has no meaning in the frequentist point of view. The frequentist doesn't allow us to measure a single event. Maybe you could say, in the Bayesian point of view, you introduce new frameworks that give you confidence intervals for a single event having not witnessed them.

Example 1.6

Consider these examples:

- "There is a 50% chance this coin will land on its 'tails' side."
- "There is a 60% chance it will rain tomorrow."
- "There is a 90% chance the cat trashed the house."

Which is frequentist? The first example, since we can repeat it multiple times.

We can also speak of the *degree of logical support* for an assertion in order to move the evaluation metric from belief to something less personal and more objective.

In the mathematical implementation of the Bayesian point of view, unknown parameters are not viewed as unknown real numbers but are replaced by random variables.

Example 1.7 (A typical statistical problem)

We have a coin that has some probability θ of landing ($H = 1$) and probability $1 - \theta$ of landing tails ($T = 0$). A coin toss is then modeled by the Bernoulli random variable X with pmf

$$f_X(k) = \begin{cases} \theta & \text{if } k = 1 \\ 1 - \theta & \text{if } k = 0 \end{cases}$$

Suppose we flip the coin 10 times and obtain the following results:

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

In the frequentist point of view, θ is a fixed by unknown real number, but we estimate it using the data. One such estimate is

$$\frac{\text{sum}}{\text{total number}} = \frac{5}{10} = 0.5$$

(Yet, in fact, the data was generated using $\theta = 0.4$).

In the Bayesian point of view, there's a high probability that θ is 0.5 but it could be something else.

We want to compute the probability density function of the parameter θ being a certain value. We'll do it in two ways, frequentist and Bayesian.

We'll view this data as the output value of 10 i.i.d. random variables

$$X_1, \ldots, X_{10}$$

defined on a sample space Ω. That is, we suppose we viewed each of these random variables at one $\omega \in \Omega$ and found

$$X_1(\omega) = 1, \quad X_2(\omega) = 0, \quad \cdots \quad X_{10}(\omega) = 1$$

Let

$$\hat{\theta} = \frac{X_1 + \cdots + X_{10}}{10}$$

which is a new random variable, called a statistic, and in this case an estimator[1] for the real number θ.

$\hat{\theta}(\omega)$ takes in some $\omega \in \Omega$. We know that

$$W = X_1 + \cdots + X_{10} \sim \mathsf{binom}[2](\mathsf{size} = 10, \mathsf{prob} = \theta)$$

Since $\hat{\theta} = \frac{W}{10}$, it takes on values $0, \frac{1}{10}, \cdots, \frac{9}{10}, 1$ and the pmf of $\hat{\theta}$ is

$$f_{\hat{\theta}}(x) = P(\hat{\theta} = x) = P(W = 10x) = \binom{10}{10x} \theta^{10x} (1-\theta)^{10-10x}$$

Why did θ come up here? $\hat{\theta}$ is conditioned on θ. Notice that the pmf of $\hat{\theta}$ depends on θ and we might emphasize this by writing $f_{\hat{\theta}}(x \mid \theta)$. Our new pmf depends on this conditional in some formulaic way.

[1] When we compute a linear regression, we compute a model (within a model class) is a maximum-likelihood estimator that reduces our loss.

[2] $\mathsf{binom}(\mathsf{size} = 10, \mathsf{prob} = \theta)$ should invoke the thought of 10 coin flips with θ probability heads each.

*** Code

When actual data is given (selecting a specific ω), we say $\hat{\theta}(\omega)$ is an estimate, but without selecting a specific ω we refer to the random variable $\hat{\theta}$ as an *estimator*.

We will, in the course, discuss properties of estimators, methods of finding them and comparing their accuracy, and apply them in analysis with hypothesis tests and by specifying confidence intervals.

In the Bayesian point of view, θ is not viewed as a fixed but unknown real number, but instead the value of a random variable Θ with some prior distribution $f_\Theta(\theta)$ for different values of θ.

The estimator $\hat{\theta}$ has conditional pmf ***

$$f_{\hat{\theta}}$$

Definition 1.8 (Prior/Posterior)

We want to build priors and posteriors. Given priors, what are posteriors (given experiment, how do we update priors?).

We start with some information about Θ expressed as a prior distribution $f_\Theta(\theta)$, and after (additional) data is collected, this is an updated to a posterior distribution using Bayes' formula. ***

§1.4 Course Logistics

We will cover

- Chapters 5-12 in Casella & Berger, *Statistical Inference* [CB02]
- Wasserman, *All of Statistics* [Was04]
- DeGroot & Schervish, *Probability and Statistics* [DS12]
- J.O. Berger, *Statistical Decision Theory and Bayesian Analysis* [Ber85]

Course is held TTh 10:30-11:50 a.m., in-person in B&H 165. Attendance is required.

TA: Liza Kolev, who will grade assignments and hold a weekly problem session.

My office hours: TBD, but for this first week please email and set up Zoom appointments.

Topics:

- ***

Weekly homework, or weekly project (a little more coding). There is no final, there is no midterm.

§2 January 31, 2022

§2.1 Quick Notes

Roughly going to move through section 5 (Casella & Berger [CB02]) through the next couple of weeks. We'll recap a bit of 1610, talk about sample mean, sample variance, and computations. Additionally, what happens when our sample mean and variance comes from normal distributions. This will be next 2 lectures. The lecture after that, we'll talk about the student-t distribution.

We'll mostly be proving results and exercise from the book.

Slides will be posted promptly, and notes will be posted after class.

Some IOUs: published Canvas page with syllabus and HW1. For planning, HW1 will be due Friday of next week, Feb 10.

§2.2

Definition 2.1 (Sample Mean)

Given random variables $X_1, \dots, X_n$ which have been observed, our sample mean is

$$\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

Definition 2.2 (Sample Variance)

Given random variables $X_1, \dots, X_n$ which have been observed, our sample variance is

$$S^2 = \frac{1}{N-1} \sum_{i=1}^{n} (X_i - \bar{X})$$

This coefficient $\frac{1}{n-1}$ plays the role in making sure that our estimator is *unbiased*.

We'll be focused on answering:

- What can we say about these estimators?
- If we know that we are taking iid samples form some distribution, say normal, what is the behavior of these estimators?

Another silly definition:

Definition 2.3 (Sample Standard Deviation)
The Sample Standard Deviation is just $S := \sqrt{S^2}$.

Theorem 2.4 (5.2.4 of [CB02])
The following:

a) $a = \hat{X}$ is the minimizer of least squares

$$\sum_{i=1}^{n}(X_i - a)^2$$

b)

$$(n-1)S^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2 = \left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2$$

Proof of a. We differentiate. We compute

$$\begin{aligned}\frac{d}{da}\left(\sum_{i=1}^{n}(X_i - a)^2\right) = \sum_{i=1}^{n} 2(X_i - a) = 0 \\ \Rightarrow \sum_{i=1}^{n}(X_i - a) = 0 \\ \Rightarrow n\bar{X} - na = 0 \\ \Rightarrow a = \bar{X}\end{aligned}$$

which is as desired. □

Proof of b. The first equality falls out of the definition. We work on the second equality, that we can prove the sum out of the summand.

$$\begin{aligned}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \sum_{i=1}^{n}(X_i^2 - \underbrace{2X_i\bar{X} + \bar{X}^2}) &= \left(\sum_{i=1}^{n} X_i^2\right) - 2\bar{X}\sum_{i=1}^{n} X_i + n\bar{X}^2 \\ &= \left(\sum_{i=1}^{n} X_i^2\right) - 2\bar{X}(n\bar{X}) + n\bar{X}^2 \\ &= \left(\sum_{i=1}^{n} X_i^2\right) - n\bar{X}^2\end{aligned}$$

which is as desired. □

We will now think of these estimators as random variables.

Theorem 2.5 (5.2.5 of [CB02])
The following:

a) That the expectation of iid random variables is the same as expectation of single random variable n times.

$$\mathsf{E}\left(\sum_{i=1}^{n} g(X_i)\right) = n\mathsf{E}(g(X_i))$$

We can simply do additivity of expectation to pull the sum out,

$$= \sum_{i=1}^{n} \mathsf{E}(g(X_i)).$$

b) The same for variance,

$$\mathsf{Var}\left(\sum_{i=1}^{n} g(X_i)\right) = n\,\mathsf{Var}(g(X_i)).$$

We use the definition of expectation here

$$\begin{aligned}
&= \mathsf{E}\left[\left(\sum_{i=1}^{n} g(X_i)]\right) - \mathsf{E}\left(\sum_{i=1}^{n} g(X_i)\right)\right]^2 \\
&= \mathsf{E}\left[\sum_{i=1}^{n} (g(X_i) - \mathsf{E}(g(X_i)))\right]^2 \\
&=^{*} \left[\sum_{i=1}^{n} \mathsf{E}\,(g(X_i) - \mathsf{E}(g(X_i)))\right]^2 \\
&= \sum_{i=1}^{n} \mathsf{Var}(g(X_i)) \\
&= n\,\mathsf{Var}\,g(X_i)
\end{aligned}$$

where $(*)$ is

$$\mathsf{E}(g(X_i) - \mathsf{E}(g(X_i)))^2 = \underbrace{\mathsf{E}(g(X_i) - \mathsf{E}(g(X_i)))}_{\mathsf{Covar}}\,\mathsf{E}(g(X_i) - \mathsf{E}(g(X_j)))$$

Notice we're still talking about these without knowledge of the underlying distributions. We'll do one more version of that.

Theorem 2.6 (5.2.6 of [CB02])

Let $X_1, \ldots, X_n$ be iid samples from a population[a] with mean μ and variance σ^2. We can say the following:

a) $\mathsf{E}\,\bar{X} = \mu$.

Since

$$\mathsf{E}(\bar{X}) = \mathsf{E}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n}\mathsf{E}(X_i) = \frac{1}{n}\cdot n\cdot\mu = \mu.$$

b) $\mathsf{Var}\,\bar{X} = \frac{\sigma^2}{n}$.

Since

$$\mathsf{Var}\,\bar{X} = \mathsf{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \mathsf{Var}\left(\sum_{i=1}^{n}\frac{1}{n}X_i\right) = n\,\mathsf{Var}\left(\frac{1}{n}X_i\right) = \frac{n}{n^2}\mathsf{Var}(X_i) = \frac{1}{n}\sigma^2.$$

c) $\mathsf{E}\,S^2 = \sigma^2$. *Proved below.*

[a] Not necessarily normal, this is weaker.

Proof of c. We have

$$\begin{aligned}
\mathsf{E}\,S^2 &= \mathsf{E}\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right)\\
&= \mathsf{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^{n}x_i^2\right) - \frac{1}{n-1}\left(\sum_{i=1}^{n}2X_i\bar{X}\right) + \frac{1}{n-1}\left(\sum_{i=1}^{n}\bar{X}^2\right)\right)\\
&= \mathsf{E}\left(\frac{1}{n-1}\left(\sum_{i=1}^{n}X_i^2\right) - \frac{n}{n-1}\bar{X}^2\right)\\
&= \frac{1}{n-1}\mathsf{E}\left(\sum_{i=1}^{n}X_i^2\right) - \frac{n}{n-1}\mathsf{E}(\bar{X}^2)\\
&= \frac{n}{n-1}\mathsf{E}(X_i^2) - \frac{n}{n-1}\mathsf{E}(\bar{X}^2)
\end{aligned}$$

Using the rule $\mathsf{E}(X^2) - \mathsf{E}(X)^2 = \mathsf{Var}(X)$,

$$\begin{aligned}
&= \frac{n}{n-1}\left(\mathsf{Var}(X_i) - \mathsf{E}(X_i)^2 - \mathsf{Var}(\bar{X}) + \mathsf{E}(\bar{X})^2\right)\\
&= \frac{n}{n-1}\left(\sigma^2 - \mu^2 - \frac{\sigma^2}{n} + \mu^2\right)\\
&= \frac{n}{n-1}\left(\frac{n-1}{n}\sigma^2\right) = \sigma^2
\end{aligned}$$

as desired. □

We still so far have made no assumptions about the underlying distributions of X_i, we now impose a normality assumption to compute $\bar{X}$ and S^2 (as in 5.3 of [CB02]), sampling from normal distribution with variance σ^2 and mean μ.

Our setup is now $X_i, \dots, X_n$ iid samples from $\mathcal{N}(\mu, \sigma^2)$.

Theorem 2.7 (5.3.1 of [CB02])
We have

a) $\bar{X}$ and S^2 are independent[a].

b) $\bar{X} \sim \mathcal{N}(\mu, \frac{\sigma^2}{n})$, where $\mathsf{E}\,\overline{X} = \mu$ and $\mathsf{Var}\, X_i = \frac{\sigma^2}{n}$.

c) $(n-1)\frac{S^2}{\sigma^2}$ has χ^2 distribution with $n-1$ degrees of freedom.

[a] Their covariance is 0, so any downstream application of these estimators can use this assumption.

Proof. Recall: that functions of independent random vectors are independent random variables.

The goal is to show that $\bar{X}$ and S^2 are functions of independent random vectors.

$$\begin{aligned} S^2 &= \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2 \\ &= \frac{1}{n-1}\left((X_1 - \bar{X})^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right) \end{aligned}$$

then with $\sum_{i=1}^{n}(X_i - \bar{X}) = 0$, ***

$$= \frac{1}{n-1}\left(\left(\sum_{i=2}^{n}(X_i - \bar{X})\right)^2 + \sum_{i=2}^{n}(X_i - \bar{X})^2\right)$$

We can now say that S^2 is a function of $X_2 - \bar{X}, \dots, X_n - \bar{X}$.

We write the pdf in terms of $X_1, \dots, X_n$, we'll do a change of variables, and factor it in the new variables, and get independence. Our new goal is to factor the pdf to show independence.

The following is the joint pdf of $X_1, \dots, X_n$,

$$F(X_i, \dots, X_n) = \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n} X_i^2}$$

(This is just the product of multiple Gaussian distributions.)

Changing coordinates, $Y_1 = \bar{X}$, $Y_2 := \bar{X} - X_2$, $Y_3 := \bar{X} - X_3$, $\dots$, $Y_n = \bar{X} - X_n$, adjusting by the Jacobian of the transformation. This transformation has Jacobian $J = \frac{1}{n}$ (by induction, writing down matrix and computing determinant).

$$F(Y_1,\dots,Y_n) = \frac{n}{(2\pi)^{n/2}} \cdot e^{-\frac{1}{2}\left(Y_1 - \sum_{i=2}^n Y_i\right)^2}$$

$$= \frac{n}{(2\pi)^{n/2}} \cdot e^{-\frac{1}{2}\sum_{i=2}^n (Y_i + Y_1)^2}$$

$$= \left[\left(\frac{n}{2\pi}\right)^{\frac{1}{2}} e^{-\frac{nY_1^2}{2}}\right] [\text{alksjdhfalskjd}^{***}]$$

in the end Y_1 independent from $(Y_2,\dots,Y_n)^2$ so $\overline{X}$ independent from S^2. □

§3 February 2, 2022

§3.1 Samples from Normal Distributions, *continued*

We have a population and we are able to compute small samples. We can compute estimators on the small samples and want to gain information from them.

Recall: last time, we were discussing what we can say about distributions from samples from normal distributions.

Theorem (theorem 2.6, 5.3.1 of [CB02])
$X_1, \dots, X_n$ are random samples from $\mathcal{N}(\mu, \sigma^2)$.

a) $\bar{X}$ and S^2 are independent random variables.

- Briefly, we expressed $X_1, \dots, X_n$ using a change of variable/transformation to $Y_1 = \bar{X}, Y_2 = \bar{X} - X_2, \dots, Y_n = \bar{X} - X_n$ and factored the pdf.

b) $\bar{X}$ has $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$ distribution.

c) $(n-1)\frac{S^2}{\sigma^2}$ has χ^2 distribution with $(n-1)$ degrees of freedom.

Proof of b. Recall: moment generating functions.

We use as fact that

$$M_{\bar{X}}(t) = \left[M_X\left(\frac{t}{n}\right)\right]^n.$$

and

$$M_X\left(\frac{t}{n}\right) = \exp\left[\frac{\mu t}{n} + \frac{\sigma^2 (t/n)^2}{2}\right]$$

Then we see that

$$M_{\bar{X}}(t) = \exp\left[\mu t + \frac{\sigma^2}{n}\frac{t^2}{2}\right]$$

which is the moment generating function of $\mathcal{N}\left(\mu, \frac{\sigma^2}{n}\right)$. □

Moving on to part c of the theorem.

χ^2 pdf:

$$f(x) = \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{P/2}} x^{\frac{p}{2}-1} e^{-x/2}$$

where Γ distribution is black-boxed. We say this distribution has p degrees of freedom.

Fact environment

Facts:

a) If Z is $\mathcal{N}(0,1)$, then $Z^2 \sim \chi^2$.

b) $X_1, \ldots, X_n$ independent, with $X_i = \chi^2_{p_i}$.

$$X_1 + \cdots + X_n \sim \chi^2_{\sum p_i}$$

Proof of theorem 2.6 c). WLOG, assume $\mu = 0, \sigma = 1$ such that $X_i \sim \mathcal{N}(0,1)$.

We induct on n. Let $\bar{X}_k, S_k^2$ be the sample mean and sample variance computed on the first k random variables.

We claim without proof that

$$(n-1)S_n^2 = (n-2)S_{n-1}^2 + \left(\frac{n-1}{n}\right)\left(X_n - \bar{X}_{n-1}\right)^2 \tag{3.1}$$

Base Case $n = 2$. We have

$$S_2^2 = \frac{1}{2}(X_2 - X_1)^2 = \left(\frac{X_2 - X_2}{\sqrt{2}}\right)^2$$

We know that[3]

$$X_2 - X_1 \sim \mathcal{N}(0,2)$$
$$\frac{X_2 - X_1}{\sqrt{2}} \sim \mathcal{N}(0,1)$$

so $S_2^2 \sim \chi_1^2$.

Inductive Hypothesis. $(k-1)\frac{S_k^2}{\sigma^2} \sim \chi^2_{k-1}$.

Inductive Step. Let $n = k+1$. Using eq. (3.1),

$$\underbrace{kS_{k+1}^2}_{\chi_k^2} = \underbrace{(k-1)S_k^2}_{\chi^2_{k-1}} + \underbrace{\left(\frac{k}{k+1}\right)\left(X_{k+1} - \bar{X}_k\right)^2}_{\chi_1^2}$$

Using the fact that $\bar{X}_k$ and S_k^2 are independent ($\bar{X}$ and S^2 are independent verbatim), we can add the χ^2's. We'll maintain the independence tagging X_{k+1} onto the proof.

We now want to show that $\sqrt{\frac{k}{k+1}}(X_{k+1} - \bar{X}_k)$ is $\mathcal{N}(0,1)$.

[3] If $X_1 \sim \mathcal{N}(\mu_1, \sigma_1^2)$ and $X_2 \sim \mathcal{N}(\mu_2, \sigma_2^2)$, then $X_2 - X_1 \sim \mathcal{N}(\mu_2 - \mu_1, \sigma_1^2 + \sigma_2^2)$

We know $\mathsf{Var}\,\bar{X}_k = \frac{1}{k}$ and $\mathsf{Var}\,X_{k+1} = 1$. Subtracting, $X_{k+1} - \bar{X}_k \sim \mathcal{N}(0, \frac{k+1}{k})$ (from above). This gives us exactly

$$\sqrt{\frac{k}{k+1}}\left(X_{k+1} - \bar{X}_k\right) \sim \mathcal{N}(0,1)$$

which was as desired.

□

We'll use this fact:

$$(n-1)\frac{S_n^2}{\sigma^2} \sim \chi^2_{n-1}$$

in deriving the Student's t-distribution.

§3.2 Student's t-distribution

Remark. William Gosset worked at the Guinness brewery, and he was measuring barley. He wanted to know whether the barley was representative of the population.

Guinness did not allow for publishing under his own name, so he published under "Student".

We have the following game: $X_1, \ldots, X_n$ are random samples from $\mathcal{N}(\mu, \sigma^2)$.

To normalize X_i, we can do

$$\frac{X_i - \mu}{\sigma} \sim \mathcal{N}(0,1).$$

To normalize $\bar{X}$, we can do

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0,1)$$

If we *know* the true variance, we can go from sample mean to true mean by using this fact.

What if we don't have the true variance? Our best bet is the sample variance. Let's look instead at

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} \sim ? \tag{3.2}$$

where S is the sample variance. What is the distribution of *this* random variable now (before, we knew it was $\mathcal{N}(0,1)$)?

Let's derive the distribution of eq. (3.2).

$$\frac{\bar{X} - \mu}{S/\sqrt{n}} = \frac{(\bar{X} - \mu)/\left(\frac{\sigma}{\sqrt{n}}\right)}{\sqrt{S^2/\sigma^2}}$$

Since we know that $(n-1)\frac{S^2}{\sigma^2} = \chi^2_{n-1}$, we know that $S/\sigma = \sqrt{\frac{\chi^2_{n-1}}{n-1}}$

$$\frac{\bar{X}-\mu}{S/\sqrt{n}} = \frac{(\bar{X}-\mu)/\left(\frac{\sigma}{\sqrt{n}}\right)}{\sqrt{\frac{\chi^2_{n-1}}{n-1}}}$$

Using $U \sim \mathcal{N}(0,1)$ and $V \sim \chi^2_p$,

$$= \frac{U}{\sqrt{V/p}}$$

Now this reduces to finding the distribution of $\frac{U}{\sqrt{V/p}}$. We get dirty with the pdfs.

$$F_{U,V}(u,v) = \frac{1}{(2\pi)^{1/2}} e^{-u^2/2} \cdot \frac{1}{\Gamma\left(\frac{p}{2}\right) 2^{p/2} v^{p/2-1} e^{-v/2}}$$

with bounds $u \in (-\infty, \infty), v \in (0, \infty)$

We do a change of variables $t = \frac{u}{\sqrt{v/p}}$ and $w = v$, with Jacobian $J = \left(\frac{w}{p}\right)^{1/2}$[4].

We now write

$$F_{T,W}(t,w) = F_{U,V}\left(t\sqrt{w/p}, w\right) \cdot \left(\frac{w}{p}\right)^{1/2}$$

We really want the distribution for T, so we evaluate

$$\begin{aligned}
\int_0^\infty F_{T,W}(t,w)\ \mathrm{d}w &= \int_0^\infty F_{U,V}\left(t\sqrt{w/p}, w\right) \cdot \left(\frac{w}{p}\right)^{1/2}\ \mathrm{d}w \\
&= \frac{1}{(2\pi)^{1/2}} \cdot \frac{1}{\Gamma\left(p/2\right) 2^{p/2}} \int_0^\infty e^{-\frac{t^2 w}{2p}} w^{p/2-1} e^{-w/2} \left(\frac{w}{p}\right)^{1/2}\ \mathrm{d}w \\
&= \frac{1}{(2\pi)^{1/2}} \cdot \frac{1}{\Gamma\left(p/2\right) 2^{p/2} p^{1/2}} \int_0^\infty \underbrace{e^{-\frac{1}{2}\left(t^2/p+1\right)} w^{\frac{p+1}{2}-1}}_{\text{pdf of } \Gamma(\cdots)}\ \mathrm{d}w
\end{aligned}$$

Which gives

$$F_T(t) = \frac{1}{(2\pi)^{1/2}} \cdot \frac{1}{\Gamma(p/2) \cdot 2^{p/2} p^{1/2}} \cdot \Gamma\left(\frac{p+1}{2}\right) \left[\frac{2}{1+t^2/p}\right]^{\frac{p+1}{2}}$$

which is the Student's t-distribution.

Question. Why do we care about the Student's t-distribution?

[4] We compute matrix $\begin{bmatrix} \frac{dt}{dv} & \end{bmatrix}$ write out

Let

$$T_n \simeq \frac{\bar{X}_n - \mu}{S_n/\sqrt{n}}$$

and we'll see that the limit of T_n as n goes to infinity approaches $\mathcal{N}(0,1)$ in distribution:

$$T_n \overset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0,1).$$

§4 February 7, 2022

A quick remark with this field of mathematical statistics. Traditionally in mathematics, we're very accustomed to building up layers of abstraction. For example, you can't really jump to the end to Galois theory with Algebra. However, this field is quite different. It *is* possible to jump to major conclusions before coming back to talk about the layers of abstraction.

We'll talk about hypothesis testing here, and then loop back to discuss the t-test.

§4.1 Hypothesis Testing

The idea, roughly, is for us to have a *null* hypothesis and *alternative* hypothesis. We want to talk about tests that allow us to do so, and the quality of those tests.

Definition 4.1 (Hypothesis Test)

We define a hypothesis test:

1. We partition parameter space Θ into Θ_0, Θ_A disjoint. For some event Θ, we say that our null hypothesis is
$$H_0 : \Theta \in \Theta_0$$
and we say that our alternative hypothesis is
$$H_A : \Theta \in \Theta_A$$

2. In our hypothesis test, our goal is to find an appropriate subset of outcomes $R \subset \mathcal{X}$ (called the rejection region). Our hypothesis test is, if $x \in R$, reject the null. Otherwise, accept the null.

Usually,
$$R = \{x \mid T(x) > c\}$$
where $T(x)$ is our estimator or statistic, and c is a critical value. Our goal is to find good Ts and cs for our hypothesis test.

Today, our estimator is always going to be the *sample mean*.

Example

Examples of hypothesis tests are p-tests, t-tests, χ^2-tests, likelihood ratio tests, etc.

There are some following outcomes of hypothesis tests[5]:

[5] And these are the only outcomes. The null is either true or not, and we make some binary decision.

	H_0 **is true**	H_A **is true**
Retain null	Wonderful!	Type II error
Reject null	Type I error	Wonderful!

A Type I error is when our null is true, yet we rejected it. A Type II error is when the alternative is true, but we kept the null.

Given a test, we want to first compute the probability of committing an error. Once we've done that, we want to ask what the *quality* of our test is.

Let's try to compute the probability of an error:

Example 4.2

Suppose we have normal distribution with $\sigma = 10$, μ unknown. We take a random sample, $n = 25$.

Hypotheses. Our null hypothesis $H_0 : \mu = 170$. Alternative hypothesis $H_A : \mu > 170$.

Test. Our test is that we reject the null if $\bar{X} \geq 172$.

Here, our estimator T is the sample mean, and our critical value is 172.

Question. What is the probability of committing a Type I error?

A Type I error is when we reject the null but the null is true. That's to say

$$\Pr(\bar{X} \geq 172) \text{ if } \mu = 170$$

We know the distribution of $\bar{X}$ for $n = 25$, so we can write

$$\begin{aligned} &= \Pr\left(\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \geq \frac{172 - \mu}{\sigma/\sqrt{n}}\right) \\ &= \Pr(Z \geq 1) \end{aligned}$$

Where $Z \sim \mathcal{N}(0, 1)$, which gives

$$= 1 - \Phi(1) \approx \boxed{.16}$$

Going back to our initial introduction, we want to compute the *quality* of our test.

Definition 4.3 (Power)

The power of a test is the probability that we reject the null when H_A is true.

§5 February 9, 2022

Rest in notebooks.

Let us have two normal distributions with mean μ_A and μ_B, $\sigma = 1$, and decision threshold $c = \frac{\mu_A+\mu_b}{2}$.

Our goal is to get the Type I and Type II error under δ. Because everything is symmetrized, Type I error equals Type II error.

Theorem 5.1

If

$$n \geq \frac{4z_\delta}{(\mu_A - \mu_B)^2} > \frac{8 \log 1/\delta}{(\mu_A - \mu_B)^2}$$

then the hypothesis test is correct with probability $\geq 1 - \delta$.

Proof. We want to bound our type II error.

$$\begin{aligned}
\Pr_B(\bar{X} \leq C) &\leq \delta \\
\Pr_B\left(\frac{\bar{X} - \mu_B}{\sigma/\sqrt{n}} \leq \frac{c - \mu_B}{\sigma/\sqrt{n}}\right) &\leq \delta \\
\Pr\left(z_\delta \leq \frac{c - \mu_B}{\sigma/\sqrt{n}}\right) &\leq \delta \\
\frac{\sigma z_\delta}{\sqrt{n}} &\leq c - \mu_B \\
c &\geq \mu_B + \frac{\sigma z_\delta}{\sqrt{n}} \\
\frac{\mu_A + \mu_B}{2} - \mu_B &\geq \frac{\sigma z_\delta}{\sqrt{n}} \\
\frac{\mu_A + \mu_B}{2} &\geq \frac{\sigma z_\delta}{\sqrt{n}} \\
\Rightarrow \sqrt{n} &\geq \frac{2\sigma z_\delta}{\mu_A - \mu_B} \\
n &\geq \frac{4z_\delta^2}{(\mu_A - \mu_B)^2}
\end{aligned}$$

□

Lemma 5.2

$z_\delta > \sqrt{2 \log \frac{1}{\delta}}$

Proof. We want to compute z_δ. We want

$$\int_z^\infty \frac{1}{\sqrt{2\pi}} e^{-x^2/2}\, dx \leq \delta$$

Let's suppose we have $X \sim \mathcal{N}(0,1)$, then

$$\Pr(X > t) = \Pr(e^{\lambda x} > e^{\lambda t}), \lambda > 0$$

Using Markov's inequality, this

$$< \frac{\mathbb{E}[e^{\lambda x}]}{e^{\lambda t}} \overset{\text{mgf}}{=} e^{\lambda^2/2 - \lambda t}$$

The minimum happens when $\lambda = t$ which is

$$\Pr(X > t) \leq e^{-t^2/2}$$

Then reintroducing z_δ,

$$\Pr(X > z_\delta) \leq e^{-z_\delta^2/2} < \delta$$

and

$$-\frac{z_\delta^2}{2} < \log \delta$$

□

Motivation: this silly hypothesis test is guaranteed a certain success rate based on n. What is the best possible bound on n over all algorithms? Is it attainable?

§6 February 14, 2022

Today we'll talk about p-values. On Thursday, we'll continue to discuss optimal samples sizes for optimal errors.

§6.1 The p-value & the p test

Definition 6.1 (p-value)

Let $W(X)$[6] be some test statistic on a raw value X, such that large values of $W(X)$ give evidence that H_1 is true. For each observed sample x, define $p(x)$ to be

$$p(x) := \sup_{\theta \in \Theta_0} \Pr_\theta(W(X) \geq w(x)).$$

$p(x)$ is called the p-value.

Example 6.2

If $\Theta_0 = \{\theta\}$, this implies that the p-value is just $\Pr_\theta(W(X) \geq w(x))$.

The p-value is the probability, under H_0, of observing a value of the test statistic that is more extreme than what we have observed.

Example 6.3

Let $W(x) = \bar{x}$. We have 2 normal distributions centered at μ_A and μ_B.

If we fixed a Type I error α, we computed a threshold on which we could achieve that Type I error c_α. Our hypothesis test was to reject the null if our sample mean was greater than or equal to c_α.

Suppose we collect n samples and compute $\bar{x} \geq c_\alpha$, we should then reject the null.

We can come to the same conclusion if we used the p-value.

$$p(\bar{x}) = \Pr_{\mu_A}(\bar{X} \geq \bar{x}) \leq \Pr_{\mu_A}(\bar{x} \geq c_\alpha)$$
$$p(\bar{x}) \leq \alpha.$$

so this suggests we could also conduct hypothesis test as follows: we reject the null if $p(\bar{x}) \leq \alpha$.

[6] In understanding this definition, we can think of $W(X)$ as the sample mean, as we've seen before.

This is called the p-test.[7]

Let's compute the probability of committing a Type I error using p-test. In other worse, we compute

$$\Pr(p(\hat{x}) \leq \alpha).$$

We'll need the distribution of p first to compute this.

Lemma 6.4

The p-value of a statistic drawn from a continuous distribution is Uniform$(0, 1)$.

Proof. We defined

$$p(x) = \Pr_{\theta_0}(W(X) \geq w(x))$$

is the cdf of $W(X)$, suffices to show that the cdf of any random variable is uniform[8]. □

Particularly, if $p(\hat{x})$ is uniform, we have

$$\Pr(p(\hat{x}) \leq \alpha) = \alpha.$$

We now have two ways of conducting a hypothesis test:

1. Critical point test, covered last week: We set α and compute threshold c_α. Reject if $\bar{x} \leq c_\alpha$ (or $\geq$).

2. p-test: We set α and compute p-value as a function of our test statistic. Reject if $p(\bar{x}) \leq \alpha$.

This also allows us to talk about the p-value of arbitrary tests.

§6.2 Wald Test

This will be our third test.

We have a scalar θ, test statistic $\hat{\theta}$ which is an estimate of θ, and standard error $\hat{\text{SE}}$ of $\hat{\theta}$.

[7] As opposed to the *critical point test* from before, selecting a critical point c.

[8] That is, choose random X, with cdf $F_X(x)$. $Z : F_X(x)$ is uniform. We compute the CDF of Z.

$$\begin{aligned} F_Z(x) &= \Pr(F_X(x) \leq x) \\ &= \Pr(x \leq F_X^{-1}(x)) \\ &= F_X(F_X^{-1}(x)) = x \end{aligned}$$

which is the cdf of uniform. (A remark on F_X^{-1}, use pseudoinverse if cdf F is not continuous).

Our null hypothesis is $H_0 : \hat{\theta} = \theta_0$. Our alternative hypothesis is $H_1 : \hat{\theta} \neq \theta_0$.

We impose the assumption on $\hat{\theta}$, which is that $\hat{\theta}$ is asymptotically normal.

$$\frac{\hat{\theta} - \theta_0}{\hat{\text{SE}}} \rightsquigarrow \mathcal{N}(0, 1)$$

The test is: Reject null if

$$|W| \geq Z_{\alpha/2} := \Phi^{-1}\left(1 - \frac{\alpha}{2}\right)$$

where

$$W = \frac{\hat{\theta} - \theta_0}{\hat{\text{SE}}}$$

Lemma 6.5

The Type I error of this test is α, asymptotically[9].

Proof. Under the null,

$$\Pr_{\theta_0}\left(|W| > Z_{\alpha/2}\right) \rightsquigarrow \Pr_{\theta_0}\left(|Z| > Z_{\alpha/2}\right) = \frac{\alpha}{2} + \frac{\alpha}{2}.$$

□

Let's compute the p-value of the Wald test. Again,

$$\Pr_{\theta_0}(|W(x)| \geq |w(x)|) \rightsquigarrow \Pr_{\theta_0}(|Z| \geq |w(x)|) = 2\Phi(-|w(x)|)$$

§6.3 t-test

Remember before that we wanted to compare if two distributions were the same. The t-statistic satisfies a t distribution which converged to a normal.

The t test frames around the t statistic. Let's say we have $X_1, \ldots, X_n$ iid from $\mathcal{N}(\mu, \sigma^2)$ where μ and σ are unknown, we want to test whether $\mu = \mu_0$ or not.

The statistic we'll use is

$$T = \frac{(\bar{X}_n - \mu_0)}{S_n/\sqrt{n}}$$

$T \sim \mathcal{N}(0, 1)$ under H_0. The t-test is: reject null when $T > t_{n-1, \frac{\alpha}{2}}$ which will give us a Type I error α.

With a large number of samples, this test converges to the Wald test and the T statistic converges to the W statistic in the Wald test.

[9] That is, take arbitrarily large number of samples.

§7 February 16, 2022

§7.1 MLEs and Likelihood Ratio Tests

Today we'll work toward proving the Neyman-Pearson. The idea is to give the *optimal* test to distinguish between two *simple* hypotheses.

By simple, we mean that we wish to distinguish whether two samples come from the same distribution, or different distributions.

By optimal, we mean that for a fixed α, we hope to minimize β.

Definition 7.1

If $X_1, \dots, X_n$ are iid samples from a population with pdf $f(x \mid \theta_1, \dots, \theta_n)$. Then the likelihood function of X is

$$L(\theta \mid X).$$

That is, we fix the samples and ask about our parameters given our fixed samples. Another way to think of this is that given some samples, what is the best model to fit to our sample. This is defined as

$$\begin{aligned} L(\theta \mid X) &= L(\theta_1, \dots, \theta_k \mid X_1, \dots, X_n) \\ &= \prod_{i=1}^{n} f(X_i \mid \theta_1, \dots, \theta_n) \end{aligned}$$

Definition 7.2 (Maximum Likelihood Estimator)

For each sample point x, let $\hat{\theta}(x)$ be a parameter value at which $L(\theta, x)$ is maximized (as a function of θ). $\hat{\theta}(x)$ is the maximum likelihood estimator of θ based on the data x.

MLE is the parameter point for which the observed sample is most likely.

If our likelihood function is differentiable in θ, then the possible candidates for MLEs are those critical points

$$\frac{\partial}{\partial \theta_i} L(\theta \mid x) = 0$$

for $i = 1, \dots, k$.

Example 7.3

Let $X_i, \dots, X_n$ be iid from $\mathcal{N}(\theta, 1)$. What is $L(\theta \mid X)$?

$$L(\theta \mid X) = \prod_{i=1}^{n} \frac{1}{(2\pi)^{1/2}} e^{-\frac{1}{2}(X_i-\theta)^2}$$
$$= \frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2}$$

What's our maximum likelihood? We set $\frac{d}{d\theta} L(\theta \mid X) = 0$. When we differentiate, the constant won't matter, the exponent won't matter, so we just want our exponent to be 0[10].

$$\sum_{i=1}^{n} (X_i - \theta) = 0$$

and solving for θ gives that $\theta = \bar{X}$.

In this setup, the best guess for θ is the sample mean.

When n goes to infinity, the sample mean will converge to the true mean.

We'll spend the next two weeks computing the maximum likelihood estimators.

Let Θ be the entire parameter space. We define

Definition 7.4 (Likelihood Ratio Test)

The likelihood ratio test statistic for testing $H_0 : \theta \in \Theta_0$ vs $H_1 : \theta \in \Theta_0^c$ is

$$\lambda(x) = \frac{\sup_{\theta \in \Theta_0} L(\theta \mid x)}{\sup_{\theta \in \Theta} L(\theta \mid x)}.$$

We should reject the null if $\lambda(x) \leq c$. That is, the probability is low that the null contains the parameter.

Example 7.5

Let $X_1, \ldots, X_n$ be iid from $\mathcal{N}(\theta, 1)$. $H_0 : \theta = \theta_0$ and $H_1 : \theta \neq \theta_0$.

$$\lambda(x) = \frac{L(\theta_0 \mid x)}{L(\bar{X} \mid x)}$$
$$= \frac{\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\theta)^2}}{\frac{1}{(2\pi)^{n/2}} e^{-\frac{1}{2}\sum_{i=1}^{n}(X_i-\bar{X})^2}}$$
$$= \exp\left[\left(\sum_{i=1}^{n} -(X_i-\theta_0)^2 + (X_i-\bar{X})^2\right)/2\right]$$

[10] Remove constants and keep on chain-ruling.

Using identity $\sum_{i=1}^n (x_i - \theta_0)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \theta_0)^2$

$$\lambda(x) = \exp\left(-\frac{n}{2}(\bar{X} - \theta_0)^2\right)$$

Then we have rejection region

$$\begin{aligned}\{x \mid \lambda(x) \le c\} &= \{x \mid \exp\left(-\frac{n}{2}(\bar{X} - \theta_0)^2\right) \le c\} \\ &= \left\{x \;\middle|\; |\bar{X} - \theta_0| \ge \sqrt{\frac{-2\log c}{n}}\right\}\end{aligned}$$

This matches up with what we saw last week—we reject the null more easily the farther away from the mean we are.

§7.2 Neyman-Pearson Lemma

The point of the Neyman-Pearson Lemma is that in testing simple hypotheses, the likelihood ratio test is the optimal test that under a Type I error, it minimizes Type II error. There is a proof ot this lemma in Casella-Berger that is unwieldly and set-theoretic.

We'll prove it for *simple* hypotheses, while the book proves it for the general likelihood ratio test.

We'll first define some notation.

Consider simple hypotheses H_0, H_1. This means we are choosing between two known distributions. In particular θ_0 is one distribution, θ_1 is another distribution. We can take a slightly different definition of the likelihood ratio test of the form

$$\text{if } \lambda(x) = \frac{p_1(x)}{p_0(x)} > \lambda$$

we accept H_1, otherwise reject null.[11]

Let R_T be the subset of X where an arbitrary test T decides H_1.

We denote

$$P_0(R_T) = \int_{R_T} p_0(x)\, dx = \Pr(\text{Type I error})$$

we'll also say that

$$P_1(R_T) = \int_{R_T} p_1(x)\, dx = \text{power}$$

and $1 - P_1(R_T)$ is our Type II error.

Let R_{LR} (for likelihood ratio) denote the subset of x where our test decides H_1. Then,

$$R_{LR}(\lambda) = \{x \mid p_1(x) > \lambda p_0(x)\}.$$

[11] We flipped the numerator and demoninator, and are accepting the alternate instead of rejecting the null.

If we set $P_0(R_{LR}(\lambda)) = \int_{R_{LR}} p_0(x)\ dx$, we choose λ such that $P_0(R_{LR}(\lambda)) = \alpha$.

We'll want to choose such an R_T that exactly matches R_{LR}.

Theorem 7.6

Consider the Likelihood Ratio Test where

$$\text{if } \frac{p_1(x)}{p_0(x)} > \lambda \implies \text{reject null}$$

Choose $\lambda = t$ such that $P_0(R_{LR}(\lambda)) = \alpha$.

There does not exist another test T with $P_0(R_T) = \alpha$ and $P_1(R_T) \geq P_1(R_{LR}(\lambda))$.

Proof. Let $R_{NP} = R_{LR}(\lambda)$[12].

Assume T is a test such that $p_0(R_T) = \alpha$.

For $R <$ range(x),

$$P_i(R) = \int_R p_i(x)\ dx = \text{probability of } x \in R \text{ under } H_i$$

Splitting this up,

$$P_i(R_{NP}) = P_i(R_{NP} \cap R_T) + P_i(R_{NP} \cap R_T^C)$$
$$P_i(R_T) = P_i(R_{NP} \cap R_T) + P_i(R_{NP}^C \cap R_T)$$

When $i = 0$, $P_0(R_{NP}) = P_0(R_T) = \alpha$, and the first terms are the same by setup, so $P_i(R_{NP} \cap R_T^C) = P_i(R_{NP}^C \cap R_T)$.

We want to show that $P_1(R_{NP}) \geq P_1(R_T)$. Suffices to show

$$P_1(R_{NP} \cap R_T^C) \geq P_1(R_{NP}^C \cap R_T)$$

Let's work on this.

$$P_1(R_{NP} \cap R_T^C) = \int_{R_{NP} \cap R_T^C} p_1(x)\ dx$$

[12] We fixed λ, so we get rid of the notation.

Because this is in the domain R_{NP}, it is larger than the threshold

$$\begin{aligned}
&\geq^{13} \lambda \int_{R_{NP} \cap R_T^C} p_0(x)\ dx \\
&= \lambda P_0(R_{NP} \cap R_T^C) \\
&= \lambda P_0(R_{NP}^C \cap R_T) \\
&= \lambda \int_{R_{NP}^C \cap R_T} p_0(x)\ dx \\
&\geq \int_{R_{NP}^C \cap R_T} p_1(x)\ dx \\
&= P_1(R_{NP}^C \cap R_T)
\end{aligned}$$

we might observe that this is essentially the Radon-Nikodym theorem. □

We'll give a short other proof.

Theorem 7.7 (Lagrange Multiplier)

If λ is a fixed nonnegaitve number $X_0(\lambda)$.

A maximizer of

$$M(x, \lambda) = f(x) - \lambda g(x)$$

then $X_0(\lambda)$ maximizees $f(x)$ over all such x such that $g(x) \leq g(X_0(\lambda))$.

Notice we are in a constraint maximization problem, minimizing Type II error (maximizing power), subject to the constraint that Type I $\leq \alpha$.

Proof sketch. We want to maximize $P_1(x)$ such that $P_0(x) \leq \alpha$. Then $M(X, \lambda) = P_1(x) - \lambda P_0(x)$.

We'll find when we apply the theorem that $X(\lambda) = \frac{P_1(x)}{P_0(x)}$ maximizes M subject to $P_0(x) \leq \alpha$. □

§8 February 23, 2022

Reviewing schedules. We hope that this is the last week of hypothesis testing. We'll discuss the KL divergence today.

§8.1 KL Divergence

Say we have $X_1, \dots, X_n$ samples from $q(x)$, iid.

Say that we have two models for $q(x)$, $p_0(x)$ and $p_1(x)$.

Having computed these, we can compute a likelihood

$$\Lambda = \prod_{i=1}^{n} \frac{p_1(x_i)}{p_0(x_i)}$$

which is called the likelihood ratio. We also have the log likelihood

$$\hat{\Lambda}_n = \frac{1}{n} \sum_{i=1}^{n} \log \frac{p_1(x_i)}{p_0(x_i)}$$

where $\hat{\Lambda}_n$

$$L_i = \log \frac{p_1(x_i)}{p_0(x_i)}$$

and we can think of our log likelihood as a sample mean over L_is,

$$\hat{\Lambda}_n = \frac{1}{n} \sum_{i=1}^{n} L_i.$$

By the law of large numbers, $\hat{\Lambda}_n \rightsquigarrow \mathsf{E}\, L_i$. Computing this,

$$\begin{aligned}
\mathsf{E}\, L_i &= \int q(x) \cdot \log \frac{p_1(x)}{p_0(x)}\, dx \\
&= \int q(x) \cdot \log \left(\frac{p_1(x)}{p_0(x)} \cdot \frac{q(x)}{q(x)} \right) \\
&= \int q(x) \left(\log \frac{q(x)}{p_0(x)} - \log \frac{q(x)}{p_1(x)} \right) dx \\
&= \int q(x) \log \frac{q(x)}{p_0(x)}\, dx - \int q(x) \log \frac{q(x)}{p_1(x)}\, dx
\end{aligned}$$

Definition 8.1 (KL Divergence)
We have that

$$\int_{-\infty}^{\infty} q(x) \log \frac{q(x)}{p_0(x)}\, dx$$

is the KL divergence of p from q. We notate this $D(q||p)$.

Then $\mathsf{E}\, L_i = D(q||p_0) - D(q||p_1)$.

So for large n, the log LRT is

$$\mathsf{E}\,\hat{\Lambda_n} \gtrless_{H_0}^{H_1} \lambda$$

which is effectively the same as

$$D(q||p_0) - D(q||p_1) \gtrless_{H_0}^{H_1} \lambda.$$

When $\lambda = 0$,

$$D(q||p_0) \gtrless_{H_0}^{H_1} D(q||p_1).$$

When LRT is equivalent to selecting the model that most closely resembles q in the context of KL.

Example 8.2

$H_0 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_0, \sigma^2)$ and $H_1 : X_1, \ldots, X_n \overset{\text{iid}}{\sim} \mathcal{N}(\mu_1, \sigma^2)$. Let's calculate the KL divergence $D(p_1||p_0)$.

We know the pdfs of these distribution, so we just calculate it.

$$\begin{aligned}
\log \frac{p_1(x)}{p_0(x)} &= \log \left(\frac{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_1)^2\right]}{\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2\sigma^2}(x-\mu_0)^2\right]} \right) \\
&= \log \left(\frac{\exp\left[-\frac{1}{2\sigma^2}(x-\mu_1)^2\right]}{\exp\left[-\frac{1}{2\sigma^2}(x-\mu_0)^2\right]} \right) \\
&= \frac{1}{2\sigma^2}\left[(x-\mu_0)^2 - (x-\mu_1)^2\right] \\
&= \frac{1}{2\sigma^2}\left[\mu_0^2 - 2\mu_0 x + 2\mu_1 x - \mu_1^2\right]
\end{aligned}$$

By definition,

$$\begin{aligned}
D(p_1||p_0) &= \int p_1(x) \log \frac{p_1(x)}{p_0(x)}\, dx \\
&= \mathsf{E}_{p_1}\left(\log \frac{p_1(x)}{p_0(x)} \right) \\
&= \mathsf{E}_{p_1}\left[\frac{1}{2\sigma^2}\left(\mu_0^2 - 2\mu_0 x + 2\mu_1 x - \mu_1^2\right) \right] \\
&= \left[\frac{1}{2\sigma^2}\left(\mu_0^2 - \mu_1^2 + 2(\mu_1 - \mu_0) \underbrace{\mathsf{E}_{p_1}(x)}_{\mu_1} \right) \right] \\
&= \frac{1}{2\sigma^2}\left[\mu_0^2 - \mu_1^2 + 2\mu_1^2 - 2\mu_0\mu_1\right] \\
&= \frac{(\mu_0 - \mu_1)^2}{2\sigma^2}
\end{aligned}$$

We want to prove that, $D(q||p) \geq 0$ and equal if $p = q$. It's *kinda* like a metric but it's not[14]. We'll

[14] It violates the triangle inequality.

use Jensen's inequality.

Theorem 8.3

If $f(x)$ is convex, then $\mathsf{E}[f(x)] \geq f(\mathsf{E}\, x)$ and equality when if f is linear.

We say f is convex if $\forall \lambda \in [0,1]$, $f(\lambda x + (1-\lambda)y) \leq \lambda(f(x)) + (1-\lambda)(f(y))$.

Proof. We have

$$\begin{aligned} D(q||p) &= \int q(x) \log \frac{q(x)}{p(x)}\, dx \\ &= \mathsf{E}_q \left[\log \frac{q(x)}{p(x)} \right] \\ &= \mathsf{E}_q \left[-\log \frac{p(x)}{q(x)} \right] \end{aligned}$$

Using the fact that $-\log(x)$ is convex, applying Jensen's

$$\begin{aligned} &\geq -\log \mathsf{E}_q \left(\frac{p(x)}{q(x)} \right) \\ &= -\log \int q(x) \cdot \frac{p(x)}{q(x)}\, dx \\ &= -\log \int p(x)\, dx \\ &= -\log 1 = 0 \end{aligned}$$

We have equality only when $\log \frac{p(x)}{q(x)}$ is linear, which happens exactly when $p = q$. □

Assume that $0 < \alpha \leq p_i(x) \leq \beta < \infty$, $i = 0, 1$. We want to bound Pr(Type I) and Pr(Type II).

$$\log \frac{\alpha}{\beta} \leq \log \frac{p_1(x_i)}{p_0(x_i)} \leq \log \frac{\beta}{\alpha}$$

$$\log \frac{\alpha}{\beta} \leq L_i \leq \log \frac{\beta}{\alpha}$$

so L_i is bounded and $\hat{\lambda}_n$ is the sum of bounded iid random variables.

Hoeffding's inequality tells us what to do when we have sums of bounded random variables.

Theorem 8.4 (Hoeffding's Inequality)
Suppose we have $Z_1, \ldots, Z_n$ iid random variables, $a \leq Z_i < b \; \forall 1 \in 1, \ldots, n$.

We have that

$$\Pr\left(\frac{1}{n}\sum_i Z_i - \mathsf{E}(Z) > \epsilon\right) \leq e^{-2n\epsilon^2/c^2}$$

and

$$\Pr\left(\mathsf{E}(Z) - \frac{1}{n}\sum_i Z_i > \epsilon\right) \leq e^{-2n\epsilon^2/c^2}$$

where $c^2 = (b-a)^2$.

Consider hypothesis test $\hat{\Lambda}_n \gtrless_{H_0}^{H_1} 0$. Then

$$\begin{aligned}\Pr(\text{Type I}) &= \Pr(\hat{\Lambda} > 0 \mid H_0) \\ &= \Pr(\hat{\Lambda}_n - \mathsf{E}[\hat{\Lambda}_n \mid H_0] > -\mathsf{E}[\hat{\Lambda}_n \mid H_0] \mid H_0)\end{aligned}$$

Let $\epsilon = -\mathsf{E}[\hat{\Lambda}_n \mid H_0]$. Let's compute ϵ.

$$\begin{aligned}\mathsf{E}_{p_0}\left[\hat{\Lambda} \mid H_0\right] &= \int p_0(x) \log \frac{p_1(x)}{p_0(x)}\, dx \\ &= -\int p_0(x) \log \frac{p_0(x)}{p_1(x)}\, dx \\ &= -D(p_0||p_1)\end{aligned}$$

Then

$$\Pr(\text{Type I}) = \Pr\left(\hat{\Lambda}_n - (-D(p_0||p_1)) > D(p_0||p_1) \mid H_0\right)$$

Applying Hoeffding's inequality,

$$\leq \exp\left[\frac{-2nD(p_0||p_1)^2}{c^2}\right]$$

where $c^2 = \left(\log\frac{\beta}{\alpha} - \log\frac{\alpha}{\beta}\right)^2$.

We can also then compute Pr(Type II).

$$\begin{aligned}\Pr(\text{Type II}) &= \Pr(\hat{\Lambda}_n < 0 | H_1) \\ &= \Pr\left(D(p_1||p_0) - \hat{\Lambda}_n > D(p_1||p_0) \mid H_1\right) \\ &\leq \exp\left[\frac{-2nD(p_1||p_0)^2}{c^2}\right]\end{aligned}$$

This gives us bounds on our Type I and Type II errors, after Neyman-Pearson gives us that the likelihood ratio test is the optimal test.