STAT 24400 - Cheat Sheet

Hye Woong Jeon

Winter 2021

Contents

1 Functions of RV **2**
- 1.1 Discrete case . . . 3
- 1.2 Continuous case . . . 3
  - 1.2.1 Special case: invertible function . . . 3
  - 1.2.2 General case: strategy . . . 4
  - 1.2.3 Uniform Distribution . . . 4

2 Joint Distributions **4**
- 2.1 Discrete Joint Distribution . . . 5
- 2.2 Continuous Joint Distribution . . . 5
- 2.3 Conditional Distribution . . . 6
- 2.4 I.I.D. draws . . . 7
- 2.5 Order Statistics . . . 7

3 Expected Value **8**

4 Variance **9**

5 Covariance and Correlation **10**

6 Conditional Expectation / Variance **12**

7 Independence **14**

8 Bayesian Statistics **15**

9 Rejection Sampling **15**

10 Frequentist Inference **16**
- 10.1 Frequentist vs. Bayesian . . . 16
- 10.2 Confidence intervals . . . 16
- 10.3 Hypothesis testing . . . 17

11 Central Limit Theorem & Applications **17**
11.1 Chi-Square Distribution . . . 19
11.2 t-Distribution . . . 20
11.3 Inferences: Sample Mean . . . 20
11.3.1 Frequentist . . . 21
11.3.2 Bayesian . . . 22
11.4 Multiple Testing Problem . . . 22

12 Parameter Estimation **23**
12.1 Making Estimators . . . 24
12.1.1 Method of Moments . . . 24
12.1.2 Maximum Likelihood Estimation . . . 24
12.2 Bayesian Inference . . . 26
12.2.1 Potential estimators & Accuracy . . . 26
12.2.2 Credible interval . . . 27

13 Hypothesis Testing **28**
13.1 Designing hypothesis tests . . . 28
13.1.1 Conventions for choosing null/alternative hypothesis . . . 28
13.1.2 Common types of rejection regions . . . 28
13.1.3 Likelihood Ratio Test (LRT) . . . 29
13.1.4 Generalized Likelihood Test . . . 29
13.2 p-values . . . 30

14 Confidence Intervals, Hypothesis Testing, & p-values **31**

15 Multinomial Data **31**

A Cheat Sheet **33**
A.1 Discrete distribution (PMF — CDF — EV — Var) . . . 33
A.2 Continuous distribution (Dens — CDF — EV — Var) . . . 33
A.3 Expectation (μ) . . . 33
A.4 Variance (σ^2) . . . 33
A.5 Covariance . . . 34
A.6 Tricks . . . 34
A.7 Analogs . . . 35

1 Functions of RV

Definition 1.1. Let X be a random variable (continuous or discrete). Define the random variable $Y = g(X)$, where g is a function from the range of X to the reals. Then g is called a **function of a random variable**.

Remark 1.2. If X is discrete, then Y will necessarily be discrete. However, if X is continuous, then Y may be discrete, continuous, or mixed (in the univariate case).

Question 1.3. *How are X and $Y = g(X)$ related to each other? In particular, how are their distributions (probability function / density function / CDF) related?*

1.1 Discrete case

The discrete case is pretty easy. Given that the PDF of X is p_X, we have that

$$\begin{aligned} p_Y(Y = y) &= P(Y = y) \\ &= P(g(x \text{ such that } g(x) = y) \\ &= \sum P(X = x), \text{ where } g(x) = y. \end{aligned}$$

Of course, the CDF can be derived from this probability function, so we omit it here.

1.2 Continuous case

1.2.1 Special case: invertible function

Remark 1.4. See Rudin problem 5.2 for reference of when the inverse function exists and is differentiable.

Suppose X is a continuous random variable, and g is differentiable and strictly monotone. Set $Y = g(X)$. Because g is strictly monotone and differentiable, g^{-1} exists and is continuous, so Y is also a continuous random variable. Hence, given the density function f_X, we have

$$f_Y(y) = \underbrace{f_X(g^{-1}(y))}_{A} \underbrace{\left| \frac{d}{dy} g^{-1}(y) \right|}_{B}.$$

The intuition is that the density function of Y at y is equal to (A) the density function of X at x such that $g(x) = y$, scaled appropriately by (B) the rate of change of g^{-1} at y.

The reason why the scaling factor B is necessary is to preserve the total area of 1 under the density curve. By the inverse function theorem, we know that the rate of change of g^{-1} at y is equal to $\frac{1}{g'(x)}$. So we're simply scaling back down the same amount that x was scaled up by g.

Proof. Assume WLOG that g is strictly increasing.

$$\begin{aligned} P(a < Y < b) &= P(g^{-1}(a) < X < g^{-1}(b)) \\ &= \int_{g^{-1}(a)}^{g^{-1}(b)} f_X(x)dx \\ &= \int_a^b f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right| dy \end{aligned}$$

By definition of density function, $f_Y(y) = f_X(g^{-1}(y)) \left| \frac{d}{dy} g^{-1}(y) \right|$. □

A function g of the form $g(X) = aX + b$ such that $a \neq 0, b \in \mathbb{R}$ follows this special case, because g is differentiable and strictly monotone. Hence, using the logic above, we can say

$$f_Y(y) = f_X\left(\frac{y-b}{a}\right) \cdot \frac{1}{|a|}.$$

1.2.2 General case: strategy

Remark 1.5. The special case above can also be derived from this general strategy.

The general strategy is the find the CDF of the new random variable Y, then differentiate it to get the density function. This means manipulating the inequalities to get something in terms of X, and subsequently using the CDF of X to get the CDF of Y.

1.2.3 Uniform Distribution

Remark 1.6. With F representing the CDF, we assume that F is strictly increasing for the below 2 theorems.

Theorem 1.7. *Let X be a continuous random variable, and let F be its CDF. Then the random variable $Y = F(X)$ is uniformly distributed on $[0, 1]$. In other words, $Y \sim$ Uniform$[0, 1]$.*

Proof. Let F_Y be the CDF of Y. Then

$$F_Y(p) = P(Y \leq p) = P(F(X) \leq p).$$

Because F is a CDF, it is continuous. Furthermore, for any $t \in (0, 1)$, there always exists a point p_1, p_2 such that $F(p_1) < t < F(p_2)$. By the intermediate value theorem, for all $p \in (0, 1)$, there exists x such that $F(x) = p$. Also F is strictly increasing by definition of CDF, so any x' such that $F(x') \leq p$ also has that $x' \leq x$. Therefore,

$$P(F(X) \leq p) = P(X \leq x) = F(x) = p.$$

The CDF of Uniform$[0, 1]$ is the identity function, and since $F_Y(p) = p$, $Y \sim$ Uniform$[0, 1]$. □

Theorem 1.8. *Let $U \sim$ Uniform$[0, 1]$, and let F^{-1} be the CDF of a particular distribution. Then $X = F^{-1}(U)$ is distributed according to the particular distribution that F defines.*

In particular, this means that we can obtain a random variable with any distribution, as long as that distribution has an invertible CDF.

Proof. $P(X \leq x) = P(F^{-1}(U) \leq x) = P(U \leq F(x)) = F(x)$. □

2 Joint Distributions

Joint distributions refer to the related probability distributions of two or more random variables defined on the same sample space.

Definition 2.1. The joint distribution of two or more random variables $X_1, ..., X_n$ defined on the same sample space refers to a characterization of their joint probabilities.

† For the remainder of this section, we only use two random variables X and Y. It is simple to extend to cases of more than two RVs.

† The comma ',' is synonymous with 'and'.

We want to find $P((X, Y) \in A)$, where A is any reasonable area in $\mathbb{R}^2$. The CDF of joint random variables X, Y is

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x \text{ and } Y \leq y) \\ &= P(X \leq x, Y \leq y). \end{aligned}$$

We can also ask about the CDF of just X (otherwise called the **marginal distribution** of X):

$$\begin{aligned} F_X(x) &= P(X \leq x) \\ &= \lim_{y \to \infty} P(X \leq x, Y \leq y) \\ &= \lim_{y \to \infty} F_{X,Y}(x, y). \end{aligned}$$

In general, the marginal distribution of X asks what happens when you keep Y unchanged, while constraining X to a certain parameter. We can also have a marginal distribution of a subset of random variables.

2.1 Discrete Joint Distribution

When all the random variables are discrete, we can also describe their distribution with a joint probability mass function $p_{X,Y}$:

$$p_{X,Y}(x, y) = P(X = x, Y = y).$$

The marginal distribution of X can be calculated with:

$$\begin{aligned} p_X(x) &= \sum_y P(X = x, Y = y) \\ &= \sum_y p(x, y). \end{aligned}$$

2.2 Continuous Joint Distribution

Definition 2.2. When (X, Y) are continuously distributed, then there exists a **joint density function** $f_{X,Y}$ that is piecewise continuous, nonnegative, and integrates to 1, such that, for any reasonable region $A \in \mathbb{R}^2$,

$$P((X, Y) \in A) = \iint_A f_{X,Y}(x, y) dy dx.$$

For the **CDF**, we have

$$\begin{aligned} F_{X,Y}(x, y) &= P(X \leq x, Y \leq y) \\ &= \int_{x=-\infty}^{x} \int_{y=-\infty}^{y} f_{X,Y}(s, t) dt ds. \end{aligned}$$

The relationship between the joint density function and the CDF is as follows:

$$f_{X,Y} = \frac{\partial^2}{\partial x \partial y} F_{X,Y}(x, y).$$

For **marginal distributions**, we have (WLOG, we choose Y):

$$\begin{aligned} F_Y(y) &= P((X, Y) \in (-\infty, \infty) \times (-\infty, y]) \\ &= \int_{y=-\infty}^{y} \int_{x=\infty}^{\infty} f(s, t) ds dt. \end{aligned}$$

Since the marginal CDF of Y is just a single variable, we can differentiate and get the following relationship

to the marginal density function of Y :

$$\begin{aligned} f_Y(y) &= \frac{d}{dy} F_Y(y) \\ &= \int_{x=-\infty}^{\infty} f(s,y)ds. \end{aligned}$$

Remark 2.3. Clearly, there's an analogue here with the discrete case. The marginal PMF of X is $\sum_y p(x,y)$, and the marginal density of X is $\int_{y=-\infty}^{\infty} f(x,t)dt$. In both cases, we "sum" over all possible values of the other random variable(s) to get the marginal PMF / density function.

Warning 2.4. For univariate distributions, if there is no probability mass at any single point in the support of the random variable, then the random variable is continuous.

But for joint distributions, we didn't define continuous distributions in the same no-single-point-mass way. Check the definition above; nowhere did we say anything about the single-point-mass thing.

In fact, it's possible to have $P((X,Y) = (x,y)) = 0$ at every $(x,y) \in A$, and **not** have a continuous distribution (i.e. there exists no density function that satisfies the conditions above). Consider the following example:

Example 2.5. (X,Y) is a point drawn uniformly at random from the unit circle $U = \{(x,y) \in \mathbb{R}^2 : ||(x,y)|| = 1\}$. Then of course, $P((X,Y) = (x,y)) = 0$ for any $(x,y) \in U$, so there is no mass at any single point. However, $P((X,Y) \in U) = 1$, but U has zero area (it's just a line in a 2D space). Hence we have

$$P((X,Y) \in U) = 1 \neq 0 = \iint_U f(x,y)dydx.$$

2.3 Conditional Distribution

Sometimes we want to ask about the distribution of X given exact knowledge of Y. For the discrete case, the conditional mass of $X \mid Y$ is

$$\begin{aligned} p_{X|Y}(x \mid y) &= P(X = x \mid Y = y) \\ &= \frac{P(X = x, Y = y)}{P(Y = y)} \\ &= \frac{p_{X,Y}(x,y)}{p_Y(y)} = \frac{p_{X,Y}(x,y)}{\sum'_x p_{X,Y}(x',y)}. \end{aligned}$$

For the continuous case, how do you define the conditional probabiltiy with $P(Y = y) = 0$ in the denominator? For this, we approximate. I won't list the details here, but we define the conditional density of $X \mid Y$ is

$$f_{X|Y}(x \mid y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}.$$

Based on the conditional distribution definitions, we can rehash the marginal distribution of Y (WLOG) as

1. (discrete) $p_Y(y) = \sum_x p_{X,Y}(x,y) = \sum_x p_{Y|X}(y \mid x)p_X(x)$, and
2. (continuous) $f_Y(y) = \int_{x=-\infty}^{\infty} f_{X,Y}(x,y) = \int_{-\infty}^{\infty} f_{Y|X}(y \mid x)f_X(x)$.

2.4 I.I.D. draws

Definition 2.6. Suppose $X_1, ..., X_n$ are independently and identically distributed (i.i.d) from a distribution with CDF F. Then since $X_1, ..., X_n$ are all independent, the joint distribution of $X_1, ..., X_n$ can be characterized by

$$F_{X_1,...,X_n}(x_1, ..., x_n) = \prod_{i=1}^{n} F_{X_i}(x_i).$$

† If $X_1, ..., X_n$ are i.i.d from a discrete distribution, then the joint PMF is the product of the individual PMFs also.

2.5 Order Statistics

Definition 2.7. Suppose we have $X_1, ..., X_n$ i.i.d from a continuous distribution with CDF F and density f. Then the order statistics are the ranked values of the random variables, denoted in the following way:

- $X_{(1)}$ denotes $\min\{X_1, ..., X_n\}$.
- $X_{(2)}$ denotes the next smallest value. ... ...
- $X_{(n)}$ denotes $\max\{X_1, ..., X_n\}$.

Warning 2.8. Just because the random variables are i.i.d **does not mean** that the order statistics are independent. Notice that $X_{(1)} < X_{(2)}$ always, so the two are not independent.

In general, if $A < B$ for some random variables A, B, then the two are not independent (think about the support argument).

Suppose for the rest of these propositions that we are dealing with random variables $X_1, ..., X_n$ that are i.i.d from a distribution described by CDF F and density f.

Proposition 2.9. *The distribution of $X_{(1)}$ is given by*

$$F_{X_{(1)}}(x) = 1 - (1 - F(x))^n, \text{ and } f_{X_{(1)}}(x) = n(1 - F(x))^{n-1} \cdot f(t).$$

Proof. For the CDF, we have

$$\begin{aligned} F_{X_{(1)}}(x) &= P(X_{(1)} \leq x) \\ &= 1 - P(X_{(1)} > x) \\ &= 1 - (P(X_1 > x, X_2 > x, ..., X_n > x)) \\ &= 1 - P(X_1 > x)P(X_2 > x)...P(X_n > x) \\ &= 1 - (1 - F(x))^n. \end{aligned}$$

The density follows by taking the derivative of the CDF. □

Proposition 2.10. *The distribution of $X_{(n)}$ is given by*

$$F_{X_{(n)}}(x) = F(x)^n, \text{ and } f_{X_{(n)}}(x) = n(F(X))^{n-1} \cdot f(x).$$

Proof. For the CDF, we have

$$\begin{aligned} F_{X_{(n)}}(x) &= P(X_{(n)} \leq x) \\ &= P(X_1 \leq x, X_2 \leq x, ..., X_n \leq x) \\ &= P(X_1 \leq x)P(X_2 \leq x)...P(X_n \leq x) \\ &= F(x)^n. \end{aligned}$$

The density follows by taking the derivative of the CDF. □

Proposition 2.11. *The distribution of $X_{(k)}$, where $1 < k < n$, is given by*

$$F_{X_{(k)}}(x) = \sum_{j=k}^{n} \binom{n}{j} [F(x)]^j [1 - F(x)]^{n-j}, \text{ and}$$

$$f_{X_{(k)}}(x) = \frac{n!}{(r-1)!(n-r)!} f(x)[F(x)]^{k-1}[1 - F(x)]^{n-k}.$$

Proof. show later. □

3 Expected Value

Definition 3.1. Let X be a random variable. Then the expected value of X (or mean of X) is the value we expect to get from X in the long run. We denote the expected value as $E(X)$, μ_X or μ.

1. If X is discrete, then
$$E(X) = \sum x \cdot p_X(x), \text{ where } p_X \text{ is the PDF of } X.$$
2. If X is continuous, then
$$E(X) = \int_{-\infty}^{+\infty} x \cdot f_X(x)dx, \text{ where } f_X \text{ is the density function of } X.$$

If the sum / integral does not exist, then the expected value also does not exist.

Proposition 3.2. *Let A be an event, and $\mathbb{1}_A$ be the indicator random variable. Then $E(\mathbb{1}_A) = P(A)$.*

Proof. We observe that $\mathbb{1}_A \sim$ Bernoulli($P(A)$). Hence,

$$E(Y) = 0 \cdot (1 - P(A)) + 1 \cdot P(A) = P(A).$$

□

Proposition 3.3 (Linearity)**.** *Let $X_1, ..., X_n$ be random variables, and let $Y = a + b_1X_1 + ... + b_nX_n$. Then*

$$E(Y) = a + b_1E(X_1) + ... + b_nE(X_n).$$

Proposition 3.4. *Let X, Y be random variables such that $X \leq Y$ almost surely. (This means that $P(X > Y) = 0$.) Then $E(X) \leq E(Y)$.*

Intuitively, if a random variable consistently spits out values that are less than another random variable, then the average of the outputs from the former RV will obviously be less than that of the latter RV.

Proposition 3.5 (Functions of RV)**.** *Let X be a random variable, and let $Y = g(X)$, where g is any function. Then*

1. *If X is discrete, then $E(Y) = \sum_x g(x) \cdot p_X(x)$.*
2. *If X is continuous, then $E(Y) = \int_{-\infty}^{\infty} g(x) \cdot f_X(x)dx$, where the limits of integration are determined by the support of X (not Y).*

You can also just find the density / CDF of Y directly and compute the $E(Y)$ directly, but the method above is usually easier.

Proposition 3.6 (Markov's Inequality)**.** *Let X be a random variable supported on $[0, \infty)$ with mean μ. Then for any $t > 0$, $P(X \geq t) \leq \frac{\mu}{t}$. Intuitively, this means that the further a number is from the mean, the less likely it is.*

Proof. Define the random variable $Y = t \cdot \mathbb{1}_{X \geq t}$. Then we have the following cases.

1. If $X \geq t$, then $Y = t \leq X$. So $Y \leq X$.
2. If $X < t$, then $Y = 0$.. Since X is supported on $[0, \infty)$, $X > 0$. So $Y \leq X$.

Hence $Y \leq X$. Therefore, $E(Y) \leq E(X) = \mu$. Now, we have

$$E(Y) = t \cdot P(Y = t) + 0 \cdot P(Y = 0) = t \cdot P(Y = t) = t \cdot P(X \geq t).$$

From this, we have $t \cdot P(X \geq t) \leq \mu$, so $P(X \geq t) \leq \frac{\mu}{t}$.

□

4 Variance

Definition 4.1. Let X be a random variable. Then the variance of X gives a measure of how spread out the values of X are from the mean of X (i.e. its expected value). The variance of X can be calculated by

$$\text{Var}(X) = E((X - \mu_X)^2).$$

The intuition is clear: $(X - \mu_X)^2$ is a function of X that gives a random variable of the distances of X's values from the mean, and we take the mean of this new random variable.

Given that Var is defined as the expected value of a function of a random variable, it is calculated by

$$\text{Var}(X) = \sum_x (x - \mu_X)^2 p_X(x) \text{ (discrete)},$$
$$\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x)dx \text{ (continuous)}.$$

Proposition 4.2. *Let X be a random variable. Then* $\text{Var}(X) = E(X^2) - E(X)^2$.

Proof. Since $\mu_X = E(X)$, we have

$$\begin{aligned}\text{Var}(X) &= E((X - \mu_X)^2) \\ &= E(X^2 - 2X\mu_X + \mu_X^2) \\ &= E(X^2) - 2\mu_X E(X) + \mu_X^2 \\ &= E(X^2) - E(X)^2.\end{aligned}$$

□

Proposition 4.3. *For any random variable X and $t > 0$,*

$$P(|X - \mu_X| \geq t) \leq \frac{\sigma_X^2}{t^2}.$$

Proof. Define $Y = (X - \mu_X)^2$. Then $E(Y) = \sigma_X^2$. By Markov's inequality, we get

$$P(|X - \mu_X| \geq t) = P(Y \geq t^2) \leq \frac{\sigma_X^2}{t^2}.$$

□

Proposition 4.4. $\mathrm{Var}(X) = 0$ *if and only if* $P(X = \mu_X) = 1$.

Proof. Intuitively, this makes sense because if all the probability weights are on the mean, then there is no spread.

Suppose $\mathrm{Var}(X) = 0$ and $P(X = \mu_X) \neq 1$. Then for any $\epsilon > 0$, $P(|X - \mu| \geq \epsilon) > 0$. But by Chebyshev's inequality, $P(|X - \mu| \geq \epsilon) = 0$, which is a contradiction.

The other direction is straightforward, just plug-n-chug. □

Proposition 4.5. *For any random variable* X *and any constants* a, b*,*

$$\mathrm{Var}(a + bX) = b^2 \mathrm{Var}(X).$$

Proof. Intuitively, if you shift the random variable by a and scale by b, then of course the spread of the shifted RV will only be affected by the scaling factor.

$$\begin{aligned}
\mathrm{Var}(a + bX) &= E((a + bX - E(a + bX))^2) \\
&= E((a + bX - (a + bE(X)))^2) \\
&= E((b(X - E(X)))^2) \\
&= b^2 E((X - E(X))^2) = b^2 \mathrm{Var}(X).
\end{aligned}$$

□

5 Covariance and Correlation

Covariance and correlation are two ways to represent how much two random variables depend on each other.

Definition 5.1. Let (X, Y) be random variables that are jointly distributed. Then

$$\mathrm{Cov}(X, Y) = E((X - \mu_X) \cdot (Y - \mu_Y)).$$

Definition 5.2. Let (X, Y) be random variables that are jointly distributed. Then

$$\mathrm{Corr}(X, Y) = \frac{\mathrm{Cov}(X, Y)}{\sigma_X \sigma_Y},$$

where σ_X and σ_Y are the standard deviations of X and Y, respectively.

Remark 5.3. For any random variable X, $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$, and $\mathrm{Corr}(X, X) = 1$.

Proposition 5.4. Corr *is always between* -1 *and* 1*. In particular,* $\mathrm{Corr}(X, Y) = \pm 1$ *if and only if* Y *is a linear transformation of* X *i.e.* $Y = aX + b$ *for some* $a, b \in \mathbb{R}, b \neq 0$.

Proof.

$$\begin{aligned}
\mathrm{Corr}(X,Y) = \pm 1 &\iff \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y} \\
&\iff \frac{E((X-\mu_X)(Y-\mu_Y))}{\sqrt{\mathrm{Var}(X)}\sqrt{\mathrm{Var}(Y)}} \\
&\iff \frac{E((X-E(X))(a+bX-(a+bE(X))))}{b\cdot \mathrm{Var}(X)} \\
&\iff \frac{b\cdot E((X-E(X))^2)}{b\cdot \mathrm{Var}(X)} \\
&\iff \frac{b\cdot \mathrm{Var}(X)}{b\cdot \mathrm{Var}(X)} = \pm 1.
\end{aligned}$$

Essentially, X and Y are completely correlated with each other if one is just a scaled and moved version of the other. □

Proposition 5.5. $\mathrm{Cov}(a+bX, a'+b'Y) = bb'\mathrm{Cov}(X,Y)$*, where* $a, a', b, b' \in \mathbb{R}$.

Proof.

$$\begin{aligned}
\mathrm{Cov}(a+bX, a'+b'Y) &= E((a+bX-\mu_{a+bX})(a'+b'Y-\mu_{a'+b'Y})) \\
&= bb'E((X-E(X))(Y-E(Y))) = bb'\mathrm{Cov}(X,Y).
\end{aligned}$$

□

Proposition 5.6. *If* $b, b' \in \mathbb{R}$ *such that they are both nonzero, then*

$$\mathrm{Corr}(a+bX, a'+b'Y) = \mathrm{Corr}(X,Y)\cdot \ \mathrm{sign}(bb').$$

Proof.

$$\begin{aligned}
\mathrm{Corr}(a+bX, a'+b'Y) &= \frac{\mathrm{Cov}(a+bX, a'+b'Y)}{\sigma_{a+bX}\sigma_{a'+b'Y}} \\
&= \frac{bb'\mathrm{Cov}(X,Y)}{|bb'|\sigma_X\sigma_Y} = \mathrm{Corr}(X,Y)\cdot \ \mathrm{sign}(bb').
\end{aligned}$$

□

Proposition 5.7. $\mathrm{Var}(X_1+...+X_n) = \sum_{i=1}^n \mathrm{Var}(X_i) + 2\sum_{i<j}\mathrm{Cov}(X_i,X_j)$.

Proof. We prove only the $n=2$ case. We have

$$\begin{aligned}
\mathrm{Var}(X+Y) &= E((X+Y-E(X)-E(Y))^2) \\
&= E((X-E(X))^2) + E((Y-E(Y))^2) + 2E((X-E(X))(Y-E(Y))) \\
&= \mathrm{Var}(X) + \mathrm{Var}(Y) + 2\mathrm{Cov}(X,Y).
\end{aligned}$$

□

Proposition 5.8. $\mathrm{Cov}(X_1+...+X_n, Y_1+...+Y_n) = \sum_{i=1}^n\sum_{j=1}^m \mathrm{Cov}(X_i,Y_j)$.

Proof.

$$\begin{aligned}\mathrm{Cov}(\sum_i X_i, \sum_j Y_j) &= E((\sum_i X_i - \sum_i E(X_i))(\sum_j Y_j - \sum_j E(Y_j))) \\ &= E(\sum_i \sum_j (X_i - E(X_i))(Y - E(Y_j))) \\ &= \sum_i \sum_j E((X_i - E(X_i))(Y - E(Y_j))) \\ &= \sum_i \sum_j \mathrm{Cov}(X_i, Y_j).\end{aligned}$$

□

Proposition 5.9. $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$.

Proof. † Notice that $\mathrm{Cov}(X, X) = \mathrm{Var}(X)$ gives the shortcut formula for variance.

$$\begin{aligned}\mathrm{Cov}(X, Y) &= E((X - E(X))(Y - E(Y))) \\ &= E(XY) - E(X)E(Y).\end{aligned}$$

□

6 Conditional Expectation / Variance

Suppose we have jointly distributed random variables (X, Y). Then the conditional expectation $E(X \mid Y = y)$ asks what the expected value of X is, given that $Y = y$.

Definition 6.1 (Conditional Expectation)**.** In the discrete case, we have

$$\begin{aligned}E(X \mid Y = y) &= \sum_x x \cdot p_{X|Y}(x \mid y), \text{ or} \\ E(g(X) \mid Y = y) &= \sum_x g(x) \cdot p_{X|Y}(x \mid y).\end{aligned}$$

In the continuous case, we have

$$\begin{aligned}E(X \mid Y = y) &= \int_x x \cdot f_{X|Y}(x \mid y)dx, \text{ or} \\ E(g(X) \mid Y = y) &= \int_x g(x) \cdot f_{X|Y}(x \mid y)dx.\end{aligned}$$

Notice the similarity between these conditional expectation definitions, and the regular expectation definitions.

Definition 6.2 (Conditional Variance)**.** We can also define conditional variance: given that $Y = y$, what is the variability among the corresponding X values?

Variance is given by

$$\begin{aligned}\mathrm{Var}(X \mid Y = y) &= E((X - E(X \mid Y = y))^2 \mid Y = y) \\ &= E(X^2 \mid Y = y) - E(X \mid Y = y)^2.\end{aligned}$$

Remark 6.3. The intuition is that we average over all of (X, Y), and remove the pairs that are not $Y = y$, and observe the average of the X values remaining.

Theorem 6.4 (Law of Total Expectation)**.** *For jointly distributed* (X, Y)*,*

$$E(Y) = E_X(E_{Y|X}(Y \mid X)).$$

Intuition. The intuition is that $E_{Y|X}(Y \mid X = x)$ is a function of X (when X changes, then the expected value of the Y's that correspond to the X values must necessarily change). Taking the expected value of all the X's gives us the expected value of Y. □

Theorem 6.5 (Tower Law for Probability)**.** *For any event* A *and any random variable* X*,*

$$P(A) = E(P(A \mid X)).$$

Proof. Intuitively, the weighted average over every probability of A given any value of X should give the probability of A itself. For the proof, consider $\mathbb{1}_A$. Then

$$\mathbb{1}_A \sim \text{Bernoulli}(P(A)).$$

Hence, it follows that $\mathbb{1}_A \mid X \sim \text{Bernoulli}(P(A \mid X))$. Therefore, by the tower law for expectation, we have

$$E(\mathbb{1}_A) = E(E(\mathbb{1}_A \mid X)) = E(P(A \mid X)).$$

□

Theorem 6.6 (Law of Total Variance)**.** *For jointly distributed* (X, Y)*, we have*

$$\text{Var}(Y) = E(\text{Var}(Y \mid X)) + \text{Var}(E(Y \mid X)).$$

In particular, this means $E(\text{Var}(Y \mid X)) \leq \text{Var}(Y)$.

Proof. First, we explain what this formula means. This formula divides variance into two summands. In particular, this means that the variability of a random variable comes from two sources:

1. $(E(\text{Var}(Y \mid X)))$ The variability of Y even if X is known, and
2. $(\text{Var}(E(Y \mid X)))$ the variability of Y as X changes.

Furthermore, $E(\text{Var}(Y \mid X)) \leq \text{Var}(Y)$ follows because variance is always nonnegative, so the second summand is always nonnegative. This inequality suggests that **on average**, the variability of Y conditioned on X is less than the variability of Y outright.

For the proof, we have

$$\begin{aligned}
\text{Var}(Y) &= E(Y^2) - E(Y)^2 \\
&= E(E(Y^2 \mid X)) - E(E(Y \mid X))^2 \\
&= E(E(Y^2 \mid X) - E(Y \mid X)^2) + E(E(Y \mid X)^2) - E(E(Y \mid X))^2 \\
&= E(\text{Var}(Y \mid X)) + \text{Var}(E(Y \mid X)).
\end{aligned}$$

□

7 Independence

Definition 7.1. Two random variables X, Y are independent if for all (x, y), $F_{X,Y}(x, y) = F_X(x)F_Y(y)$. We denote independent random variables as $X \perp Y$.

Proposition 7.2. *In both discrete and continuous cases, $X \perp Y$ is equivalent to, for all x, y,*

1. $F_{X,Y}(x, y) = \{$*some function of* $X\} \cdot \{$*some function of* $Y\}$
2. $P(X \in A, Y \in B) = P(X \in A)P(Y \in B)$.

If $X \perp Y$ are independent, then $\{$possible (X, Y) values$\} = \{$possible X values$\} \times \{$possible Y values.$\}$ (Use the contrapositive to show non-independence.)

Proposition 7.3. *In the discrete case, $X \perp Y$ is also equivalent to, for all x, y,*

1. $p_{X,Y}(x, y) = p_X(x)p_Y(y)$
2. $p_{X,Y}(x, y) = ($*some function of* $X) \cdot ($*some function of* $Y)$.

For conditional distribution, $X \perp Y$ is equivalent to, for all x, y, $p_{X|Y}(x \mid y) = p_X(x)$.

Proposition 7.4. *In the continuous case, $X \perp Y$ is also equivalent to, for all x, y,*

1. $f_{X,Y}(x, y) = f_X(x)f_Y(y)$
2. $f_{X,Y}(x, y) = ($*some function of* $X) \cdot ($*some function of* $Y)$.

For conditional distribution, $X \perp Y$ is equivalent to, for all x, y, $f_{X|Y}(x \mid y) = f_X(x)$.

Proposition 7.5. *If $X \perp Y$, then $E(X \cdot Y) = E(X) \cdot E(Y)$. In particular, if $X \perp Y$, then for any functions g, h, $g(X) \perp h(Y)$, so $E(g(X) \cdot h(Y)) = E(g(X)) \cdot E(h(Y))$.*

Proof.

$$\begin{aligned} E(X \cdot Y) &= \int_x \int_y x \cdot y \cdot f_{X,Y}(x, y) dy dx \\ &= \int_x \int_y (x \cdot f_X(x))(y \cdot f_Y(y)) dy dx \\ &= \left(\int_x (x \cdot f_X(x)) dx \cdot \right) \cdot \left(\int_y (y \cdot f_Y(y)) dy \right) \\ &= E(X) \cdot E(Y). \end{aligned}$$

□

Proposition 7.6. *If $X \perp Y$, then* $\mathrm{Cov}(X, Y) = \mathrm{Corr}(X, Y) = 0$.

Proof. Since $\mathrm{Cov}(X, Y) = E(XY) - E(X)E(Y)$, and $E(X)E(Y) = E(XY)$ because $X \perp Y$, it follows that $\mathrm{Cov}(X, Y) = 0$. Therefore, $\mathrm{Corr}(X, Y) = 0$.

This also makes intuitive sense because two independent variables should have nothing to do with each other, which means that they are not correlated. □

Proposition 7.7. *If $X_1, ..., X_n$ are mutually independent, then*

$$\mathrm{Var}(X_1 + ... + X_n) = \mathrm{Var}(X_1) + ... + \mathrm{Var}(X_n).$$

Proof. This follows immediately from Proposition 5.7 and Proposition 7.6. □

8 Bayesian Statistics

In many statistical inference settings, we observe some random data and try to learn something about the underlying parameters. However, in Bayesian statistics, we assume that the underlying parameters themselves are random variables i.e. sampled from a distribution.

Definition 8.1 (Bayesian statistics definitions)**.** We will refer to the observed random variable as X and its underlying parameter as λ.

1. Prior distribution: the distribution that we believe the underlying parameter is sampled from.
2. Posterior distribution: the conditional distribution of λ given X.
3. Posterior mean: the conditional expected value of λ given X.
4. Posterior mode: the mode of the conditional distribution (a.k.a. posterior distribution).
5. Likelihood: the conditional distribution of X given λ.

9 Rejection Sampling

Sampling from distributions can be difficult. But here are some of the ways that we know of to sample from distributions:

1. Naturally occurring processes e.g. coin flipping model coincides with binomial distribution
2. Taking a discrete PMF p, and a random number generator $U \sim \text{Uniform}[0,1]$. (basically a partition of $[0,1]$ by probabilities)
   - If $U \leq p(x_1)$, then $X = x_1$.
   - If $p(x_1) < U \leq p(x_2)$, then $X = x_2$... and so on.
3. If a closed form CDF F is known, then taking $U \sim \text{Uniform}[0,1]$, and setting $X = F^{-1}(U)$.

But how can you sample from a distribution that isn't easy to sample from, or you don't know how to sample from it? The answer is rejection sampling.

Theorem 9.1. *Let $q(x)$ be a PMF/PDF that we know how to sample from. Suppose we want to sample from a distribution with the PMF/PDF*

$$p_*(x) = (\text{normalizing constant}) \cdot h(x).$$

Also, set $C = \max_{x \in \text{support}(h)} \frac{h(x)}{q(x)}$ (or any upper bound for that matter). Then sampling from the hierarchical model

$$\begin{cases} X \sim (\text{distribution with PDF/PMF } q) \\ A \mid X \sim \text{Bernoulli}\left(\frac{h(x)}{C \cdot q(x)}\right) \end{cases},$$

and selecting only those X values for which $A = 1$, the resulting X values will follow the distribution defined by p_*. Furthermore, C tells about the efficiency of the sampling, i.e. how often X values are "accepted".

Proof. We show for the discrete case only. The continuous case follows similarly by using the ϵ argument. Since we only accept the X values for which $A = 1$, consider $P(X = x \mid A = 1)$. We have

$$\begin{aligned} P(X = x \mid A = 1) &= \frac{P(A = 1 \mid X = x)P(X = x)}{P(A = 1)} \\ &= \frac{\frac{h(x)}{C \cdot q(x)} \cdot q(x)}{\sum_{x'} \frac{h(x')}{C \cdot q(x')} \cdot q(x')} \\ &= (\text{normalizing constant}) \cdot h(x) = p_*(x). \end{aligned}$$

Furthermore, C tells about the efficiency of sampling. We have

$$P(A = 1) = \sum_{x'} \frac{h(x')}{C \cdot q(x')} \cdot q(x') = \frac{1}{C} \sum_{x'} h(x').$$

For low values of C, the sampling efficiency is high, while for high values of C, the sampling efficiency is low. This makes sense because C tells us how large the discrepancy between our "target" distribution, and the distribution that we know how to sample from is. □

10 Frequentist Inference

10.1 Frequentist vs. Bayesian

The scenario is that we have random variable X, with X dependent on some parameter θ.

1. (Frequentist) The assumption is that θ is fixed.
2. (Bayesian) The assumption is that θ is random and sampled from a prior distribution.

10.2 Confidence intervals

Question 10.1. *Which values of θ are plausible in light of the data?*

The method is as follows. After observing X, consider a range [lower bound(X), upper bound(X)] such that for any possible θ,

$$P(\text{lower bound}(X) \leq \theta \leq \text{upper bound}(X)) \geq 1 - \alpha, \text{ where } \alpha = \text{ tolerate rate.}$$

Warning 10.2. Although θ is **not** a random variable, this probability is well-defined because the probability is actually based on X. In other words, the range defined does **not** give the probability that θ falls inside the range. Rather, the probability refers to the reliability of the estimation procedure.

Remark 10.3. You can also manipulate the inequality above to make it an inequality about X assuming that θ is a particular value. In other words, something like

$$P(\text{function}_1(\theta) \leq X \leq \text{function}_2(\theta)) \geq 1 - \alpha.$$

This means that you have that the probability that X lies in between two values **assuming the value of** θ is greater than $1 - \alpha$, which gives the "confidence level" in our estimate of θ.

Remark 10.4. Suppose 1000 researchers run the same experiment, and each generates a $1 - \alpha$ confidence interval $X_i \pm \epsilon_i$. Then since the confidence level gives a probability on the **estimation** procedure, $1 - \alpha$ of the 1000 intervals generated will contain θ. But whether or not μ falls in the interval is unknown to every researcher.

10.3 Hypothesis testing

Question 10.5. *For a particular value of θ, is this value plausible in light of the data?*

The method is as follows. We are interested in the following hypotheses:

1. Null hypothesis: $\theta = \theta_0$
2. Alternative hypothesis: $\theta \neq \theta_0$ (or $\theta > \theta_0$, or $\theta < \theta_0$)

The goal is to disprove the null hypothesis. Assuming that the null hypothesis is true i.e. $\theta = \theta_0$, construct a range of values R such that

$$P(X \in R) \geq 1 - \alpha, \text{ where } \alpha = \text{ tolerate rate.}$$

Note that P is the probability over X parameterized by θ_0, as we assumed that the null hypothesis is true. Then the decision making process is as follows: if after observing X

1. If we observe a value that falls in R, then we do not reject the null hypothesis.
2. If we observe a value that does not fall in R, then we reject the null hypothesis.

From the decision making procedure, we have

$$P(X \notin R) = 1 - P(X \in R) \leq \alpha.$$

Since an observed value that does not fall in R immediately leads to a rejection of the null, and R was based on the assumption that $\theta = \theta_0$, we have a $\leq \alpha$ chance of incorrectly rejecting the null hypothesis i.e. concluding that $\theta \neq \theta_0$ when θ is actually equal to θ_0.

11 Central Limit Theorem & Applications

Definition 11.1. Let $X_1, X_2, ...$ be i.i.d from a distribution with mean μ and variance σ^2. Then we define

1. Sample sum $S_n = X_1 + ... + X_n$, and $E(S_n) = n\mu$ and $\text{Var}(S_n) = n\sigma^2$.
2. Sample mean $\bar{X} = \frac{S_n}{n}$, and $E(\bar{X}) = \mu$ and $\text{Var}(\bar{X}) = \frac{\sigma^2}{n}$.
3. Sample variance $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, and $E(S^2) = \sigma^2$.

Proposition 11.2. *The sample mean and sample variance are unbiased estimators i.e. $E(\bar{X}) = \mu$ and $E(S^2) = \sigma^2$.*

Proof. For the sample mean, we have

$$E(\bar{X}) = \frac{1}{n}E(X_1 + ... + X_n) = \frac{1}{n}(E(X_1) + ... + E(X_n)) = \mu.$$

For the sample variance, we have

$$\begin{aligned}
E(S^2) &= E\left(\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E(((X_i-\mu)-(\bar{X}-\mu))^2\right) \\
&= \frac{1}{n-1}\left(\sum_{i=1}^{n}E((X_i-\mu)^2)\right) - \frac{n}{n-1}E((\bar{X}-\mu)^2) \text{ (because cross covariance is zero)} \\
&= \frac{n}{n-1}\sigma^2 - \frac{n}{n-1}\cdot\frac{\sigma^2}{n} = \sigma^2.
\end{aligned}$$

□

Theorem 11.3 (Central Limit Theorem)**.** *If $X_1, X_2, ...$ be i.i.d from a (reasonable) distribution, then for sufficiently large n (usually $n > 30$ suffices),*

$$\left(\text{Distr. of } Z_n = \frac{S_n - n\mu}{\sigma\sqrt{n}}\right) \approx N(0,1).$$

This is to say that

$$\lim_{n\to\infty} P(Z_n \le x) \approx \Phi(x), \text{ where } \Phi \text{ is the CDF of } N(0,1).$$

Furthermore, since $S_n, \bar{X}$, and $\mathbb{Z}_n$ are linear transformations of each other, we get

$$(\textit{Distr. of } S_n) \approx N(n\mu, n\sigma^2), \text{ and } (\textit{Distr. of } \bar{X}) \approx N\left(\mu, \frac{\sigma^2}{n}\right).$$

Proposition 11.4 (Standardization)**.** *Let $Z \sim N(\mu, \sigma^2)$. Then $\frac{Z-\mu}{\sigma} \sim N(0,1)$.*

Proof. Linear transformations of normal distributed random variables are also normally distributed. More specifically, $aZ + b \sim N(a\mu + b, a^2\sigma^2)$. Hence we have

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

□

The CLT, combined with the standardization property, gives a powerful tool. Consider the following example.

Example 11.5. Let $X \sim \text{Binom}(n,p)$. Then $X = X_1 + ... + X_n$, where $X_i \sim \text{Bernoulli}(p)$ indicates success on the ith trial. We can use the CLT and standardization to give an estimate for the probability, assuming of course that n is sufficiently large.

What is the probability that we get no more than 10 heads i.e. $P(X \le 10)$? We know that $E(X_i) = p$ and $\text{Var}(X_i) = p(1-p)$. By the CLT, X is approximately normally distributed (X is the sample sum here). By standardization, we have

$$P(X \le 10) = P\left(\frac{X-np}{\sqrt{np(1-p)}} \le \frac{10-np}{\sqrt{np(1-p)}}\right) \approx \Phi\left(\frac{10-np}{\sqrt{np(1-p)}}\right).$$

Theorem 11.6. *Let $X \sim N(\mu_1, \sigma_1^2)$, and $Y \sim N(\mu_2, \sigma_2^2)$, and suppose $X \perp Y$. Then $X + Y \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$.*

Combined with the CLT, this means if $X \perp Y$ and $X, Y \approx$ (normally distributed), then $X+Y \approx$ (normally distributed) also.

Here is the motivating theorem:

Theorem 11.7. *If $X_1, ..., X_n$ are i.i.d. from $N(\mu, \sigma^2)$, then*

1. $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$
2. $\frac{n-1}{\sigma^2} \cdot S^2 \sim \chi^2_{n-1}$
3. $\frac{\bar{X}-\mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}$.
4. $\bar{X} \perp S^2$.

Furthermore, if $X_1, ..., X_n$ are i.i.d. from any distribution with mean μ and variance σ^2, the above statements hold approximately for the appropriate random variables.

Essentially, the point is that you can construct random variables that approximately follow certain distributions.

11.1 Chi-Square Distribution

In the section above, we asked how accurately $\bar{X}$ can estimate μ. Here, we ask how accurately other random variables can estimate other information, and how these random variables are distributed.

Definition 11.8 (Chi-Square Distribution)**.** Let $Z_1, ..., Z_n$ be i.i.d. from $N(0, 1)$. Define $V = Z_1^2 + ... + Z_n^2$. Then $V \sim \chi^2_n$. The density of χ^2_n is

$$f(x) = \frac{1}{2^{\frac{n}{2}}\Gamma\left(\frac{n}{2}\right)} x^{\frac{n}{2}-1} e^{-\frac{x}{2}} \text{ for } x \geq 0.$$

Also, $E(V) = n$, and $\text{Var}(V) = 2n$. Note that $\chi^2_n = \text{Gamma}(\frac{n}{2}, \frac{1}{2})$.

Proposition 11.9. *If $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$, then*

$$\sum_{i=1}^{n} \left(\frac{X_i - \mu}{\sigma}\right)^2 \sim \chi^2_n.$$

Proof. Standardize all X_i's and use the definition of the chi-square distribution. □

Proposition 11.10. *If $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$, then*

$$\sum_{i=1}^{n} \left(\frac{X_i - \bar{X}}{\sigma}\right)^2 \sim \chi^2_{n-1}.$$

This is equivalent to saying

$$\frac{n-1}{\sigma^2} \cdot S^2 \sim \chi^2_{n-1}.$$

This means that an appropriately scaled sample variance is chi-square distributed with $n-1$ degrees of freedom.

Proof. We prove only for the case $n = 2$. We have

$$\frac{1}{\sigma^2}((X_1 - \bar{X})^2 + (X_2 - \bar{X})^2 = \frac{1}{\sigma^2}\left(\left(\frac{X_1 - X_2}{2}\right)^2 + \left(\frac{X_2 - X_1}{2}\right)^2\right)$$
$$= \left(\frac{X_1 - X_2}{\sqrt{2\sigma^2}}\right)^2.$$

Since X_1, X_2 are i.i.d. from $N(\mu, \sigma^2)$, $X_1 - X_2 \sim N(0, 2\sigma^2)$. Therefore, by standardizing, we have

$$\frac{n-1}{\sigma^2} \cdot S^2 = \left(\frac{X_1 - X_2}{\sqrt{2\sigma^2}}\right)^2 \sim \chi_1^2.$$

See proposition right before this one. □

11.2 t-Distribution

Definition 11.11 (t-distribution)**.** Let $Z \sim N(0,1)$, and $V \sim \chi_n^2$, and $Z \perp V$. (Note that V can be constructed by taking i.i.d. standard normal RVs). Then

$$T = \frac{Z}{\sqrt{\frac{V}{n}}} \sim t_n.$$

The density of t_n is

$$f(x) = (\text{normalizing constant}) \cdot \left(1 + \frac{x^2}{n}\right)^{-\frac{n+1}{2}} \text{ for } x \in \mathbb{R}.$$

Proposition 11.12. *For small n, t distribution has heavy tails i.e. $P(T \geq x)$ is much larger than $1 - \Phi(x)$. Furthermore, as $n \to \infty$, $t_n \to N(0,1)$.*

Proposition 11.13. *Let $X_1, ..., X_n$ be i.i.d. from $N(\mu, \sigma^2)$. Then*

$$T = \frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}.$$

In general, the χ^2 distribution and t^2 distribution are concerned with how the distributions of the sample mean / sample variance can provide information about the actual mean / variance. First, we explore what information we can find about the mean through the sample mean.

11.3 Inferences: Sample Mean

Question 11.14. *How accurately can $\bar{X}$ estimate μ?*

This question can be represented in two ways:

1. $P(|\bar{X} - \mu| > (\text{smth. with limit } 0)) \leq \epsilon$?

Let $\epsilon > 0$. Then let $z = \Phi^{-1}(1 - \frac{\epsilon}{2})$. Then we have

$$\begin{aligned} P(|\bar{X} - \mu| > z \cdot \frac{\sigma}{\sqrt{n}}) &= P(\bar{X} - \mu > z \cdot \frac{\sigma}{\sqrt{n}}) + P(\bar{X} - \mu < -z \cdot \frac{\sigma}{\sqrt{n}}) \\ &= P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} > z) + P(\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}} < -z) \\ &\approx \text{(by CLT) } 2(1 - \Phi(z)) \\ &= 2(1 - \Phi(\Phi^{-1}(1 - \frac{\epsilon}{2}))) = \epsilon. \end{aligned}$$

Hence, as $\epsilon \to 0$, $\bar{X}$ becomes an apt approximation of μ.

2. $P(|\bar{X} - \mu| > \epsilon) \leq$ (smth. with limit 0)?

   Unfortunately, the CLT cannot be applied in this step in the case that $\bar{X}$ is NOT normal. The CLT can only be applied when the parameter in the probability statement is fixed.

   However, we **can** estimate this with Chebyshev's inequality. We have

$$P(|\bar{X} - \mu| > \epsilon) \leq \frac{\text{Var}(\bar{X})}{\epsilon^2} = \frac{\sigma^2}{n\epsilon^2}.$$

   This scales $O(\frac{1}{n})$, so hence we obtain the same conclusion as before.

11.3.1 Frequentist

The answer to the question above can be used to construct confidence intervals. However, the answer will change depending on whether or not the true variance is known - observe that the $(1 - \epsilon)$ confidence interval above depends on σ, which may not be known.

Theorem 11.15. *Suppose $X_1, ..., X_n$ are i.i.d. from $N(\mu, \sigma^2)$. Let the tolerance / error rate be $\alpha > 0$. Then*

1. *If σ^2 is known, then let $z_{\alpha/2} = \Phi^{-1}(1 - \frac{\alpha}{2})$. Then $\bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}$ is a $(1 - \alpha)$ confidence interval for μ, or*

$$P(\mu \in \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}}) = 1 - \alpha.$$

2. *If σ^2 is unknown, then suppose $F^{-1}_{t_{n-1}}$ is the inverse CDF of t_{n-1}. Let $t_{\alpha/2} = F^{-1}_{t_{n-1}}(1 - \alpha/2)$. Then $\bar{X} \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}$ is a $(1 - \alpha)$ confidence interval for μ, or*

$$P(\mu \in \bar{X} \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{n}}) = 1 - \alpha.$$

The same results hold approximately if $X_1, ..., X_n$ are i.i.d. from any distribution with mean μ and variance σ^2.

Proof. The proof for when σ^2 is known is the same as the question above. If σ^2 is unknown, use the same method from noticing that

$$\frac{\bar{X} - \mu}{\frac{S}{\sqrt{n}}} \sim t_{n-1}, \text{ as shown in a proposition above.}$$

□

11.3.2 Bayesian

Definition 11.16. Let the tolerate / error rate be $\alpha > 0$. After observing X, construct a range R such that for the parameter θ,

$$P(\theta \in R \mid X) = 1 - \alpha.$$

Then R is the $1-\alpha$ credible interval. Note that this probability is defined in the domain of the posterior probability.

Remark 11.17. The credible interval differs from the confidence interval because the former gives the literal probability that the parameter θ falls within the interval, while the confidence interval gives the **confidence** that θ falls within the interval.

Theorem 11.18. *Let the tolerance / error rate be $\alpha > 0$. Then*

1. *If σ^2 is known, then based on a μ_0 and σ_0^2 of our choosing, define the hierarchical model*

$$\begin{cases} \mu & \sim N(\mu_0, \sigma_0^2) \\ X_1, ..., X_n \mid \mu & \sim_{i.i.d} N(\mu, \sigma^2) \end{cases}.$$

   Then as $n \to \infty$, the $1-\alpha$ credible interval approaches

$$P(\mu \in \bar{X} \pm z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} \mid X_1, ..., X_n) = 1 - \alpha.$$

   This interval aligns with the frequentist $1-\alpha$ confidence interval.

2. *If σ^2 is unknown, then based on a τ, λ, and μ_0 of our choosing, define the hierarchical model*

$$\begin{cases} (1/\sigma^2) & \sim \text{Gamma}(\tau, \lambda) \\ \mu \mid \sigma^2 & \sim N(\mu_0, \sigma^2) \\ X_1, ..., X_n \mid \mu, \sigma^2 & \sim_{i.i.d.} N(\mu, \sigma^2) \end{cases}.$$

   Then as $n \to \infty$, the $1-\alpha$ credible interval approaches

$$P(\mu \in \bar{X} \pm t_{\alpha/2} \cdot \frac{S}{\sqrt{n}} \mid X_1, ..., X_n) = 1 - \alpha.$$

   This interval aligns with the frequentist $1-\alpha$ confidence interval.

Proof. We will not go into detail. On the surface level, when σ^2 is known, the posterior distribution ends up being a normal distribution where for large enough n,

$$N(\approx \bar{X}, \approx \frac{\sigma^2}{n}).$$

Then follow the same logic as in the confidence interval case.

The logic is the same for the σ^2 unknown case, except that the posterior distribution ends up being a $t_{2\tau+n}$ distribution. □

11.4 Multiple Testing Problem

The multiple testing problem is making decisions on questions or conditions on your experiment **based on** inferences you have already made.

12 Parameter Estimation

We saw above how $\bar{X}$ and S^2 can be estimates of true μ and true σ^2, and how you can substitute them in to generate confidence / credible intervals. But what about the general case? Can you make estimates of any kind of parameter - not just for μ or σ^2?

Notation 12.1. Let $X_1, ..., X_n \sim_{i.i.d.}$ from some Distribution(θ), where $\theta \in \Theta \subset \mathbb{R}^n$ is unknown. We denote

$$f(x \mid \theta) = \text{Distribution}(\theta).$$

For example, Exponential(λ) $= f(x \mid \lambda)$. In other words, each of the X_i's are sampled from $f(x \mid \theta)$.

Remark 12.2. If in frequentist framework, then $f(x \mid \theta)$ does not represent a conditional probability, since θ is assumed to be fixed. But in Bayesian framework, $f(x \mid \theta)$ literally represents a conditional probability.

Definition 12.3 (Family of densities)**.** Suppose $X_1, ..., X_n \sim_{i.i.d.} f(\cdot \mid \theta)$. Then $f(\cdot \mid \theta)$ is called the family of densities.

Definition 12.4 (Estimator)**.** Suppose $X_1, ..., X_n \sim_{i.i.d.} f(\cdot \mid \theta)$. Let $\hat{\theta} : \mathbb{R}^n \to \Theta$, where $\hat{\theta}$ maps the data $(X_1, ..., X_n)$ to an estimate of the parameter θ. We call $\hat{\theta}$ an estimator of θ.

We say $\hat{\theta}$ is unbiased if $E(\hat{\theta}) = \theta$. For example, $\bar{X}$ and S^2 are unbiased estimators of μ and σ^2, respectively.

Notice that $\hat{\theta}$ has no dependence on the true parameter θ; $\hat{\theta}$ only has the data $X_1, ..., X_n$ as input.

Definition 12.5 (Sampling Distribution)**.** We call the distribution of $\hat{\theta}$ the sampling distribution.

Recall that the output of $\hat{\theta}$ varies depending on the inputs $(X_1, ..., X_n)$, which are i.i.d. from $f(\cdot \mid \theta)$. Hence, $\hat{\theta}$ is also a random variable, and has a distribution that actually depends on θ, unlike the estimator function itself. The true parameter θ is fixed, and $(X_1, ..., X_n)$ are generated depending on the fixed θ.

Definition 12.6 (Standard Error)**.** Any estimate of the standard deviation of $\hat{\theta}$, which is distributed according to the sample distribution, is called the standard error.

The true standard deviation of $\hat{\theta}$ may depend on θ, but the standard error need not depend on θ.

Definition 12.7 (Mean Squared Error (MSE))**.** Let $\theta \in \Theta \subset \mathbb{R}$, and let $\hat{\theta}$ be an estimator of θ. Then fixing θ – so that we have a sampling distribution – we call

$$\text{MSE} = E((\hat{\theta} - \theta)^2).$$

The MSE is "how good of an estimate $\hat{\theta}$ is of θ".

Definition 12.8 (Bias)**.** We call $(E(\hat{\theta}) - \theta)^2$ the bias.

Proposition 12.9.

1. *If $\hat{\theta}$ is unbiased, then* $\text{MSE} = \text{Var}(\hat{\theta})$.
2. *If $\hat{\theta}$ is biased, then* $\text{MSE} = (E(\hat{\theta}) - \theta)^2 + \text{Var}(\hat{\theta})$.

Proof.

1. Since $E(\hat{\theta}) = \theta$, we see that $\text{MSE} = E((\hat{\theta} - \theta)^2) = E((\hat{\theta} - E(\hat{\theta}))^2) = \text{Var}(\hat{\theta})$.
2. We have that

$$\begin{aligned}
\text{MSE} &= E((\hat{\theta} - \theta)^2) \\
&= E(((E(\hat{\theta}) - \theta) + (\hat{\theta} - E(\hat{\theta})))^2) \\
&= (E(\hat{\theta}) - \theta)^2 + E((\hat{\theta} - E(\hat{\theta}))^2) \\
&= (E(\hat{\theta}) - \theta)^2 + \text{Var}(\hat{\theta}).
\end{aligned}$$

□

Remark 12.10 (Bias / Variance tradeoff). The previous proposition tells that the magnitude of the MSE comes from two sources – the bias and the variance of $\hat{\theta}$. Usually, one is obtained at the expense of another – low bias / high variance, or high bias / low variance.

If an estimator is unbiased (0 bias), but has high variance, then clearly it isn't that great of an estimate of θ. The same is true for the other case.

12.1 Making Estimators

How can you make reasonably good estimators?

12.1.1 Method of Moments

Suppose $X_1, ..., X_n \sim_{i.i.d.} f(\cdot \mid \theta)$. Let $X \sim f(\cdot \mid \theta)$.

- If $\theta \in \mathbb{R}$, then
  1. Compute $E(X)$ as a function of θ.
  2. Compute the sample mean $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$.
  3. Solve the equation $E(X) = \bar{X}$ for θ, and set this $\theta = \hat{\theta}$.
- If $\theta \in \mathbb{R}^k$, then
  1. Compute $E(X), E(X^2), ..., E(X^k)$ as functions of θ.
  2. Compute $\frac{1}{n}\sum_{i=1}^{n} X_i, \frac{1}{n}\sum_{i=1}^{n} X_i^2, ..., \frac{1}{n}\sum_{i=1}^{n} X_i^k$.
  3. Solve the equations

$$E(X) = \frac{1}{n}\sum_{i=1}^{n} X_i,\ E(X^2) = \frac{1}{n}\sum_{i=1}^{n} X_i^2, ...,\ E(X^k) = \frac{1}{n}\sum_{i=1}^{n} X_i^k$$

  for θ, and set this $\theta = \hat{\theta}$.

12.1.2 Maximum Likelihood Estimation

Notation 12.11. We will denote the true parameter value as θ_0. Any other possible parameter value will be denoted as θ.

Definition 12.12. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} f(\cdot \mid \theta)$. Let $X \sim f(\cdot \mid \theta)$. Then the likelihood is the joint density of $(X_1, ..., X_n)$ as a function of θ, or

$$\prod_{i=1}^{n} f(X_i \mid \theta).$$

Similarly, the log likelihood is the log of the likelihood, or

$$\sum_{i=1}^{n} \log(f(X_i \mid \theta)).$$

Theorem 12.13 (Maximum Likelihood Estimation (MLE))**.** *Set $\hat{\theta}$ equal to the value of θ that maximizes the likelihood (or the log likelihood), i.e.*

$$\hat{\theta} = \text{argmax}_{\theta\in\Theta} \prod_{i=1}^{n} f(X_i \mid \theta), \text{ or } \hat{\theta} = \text{argmax}_{\theta\in\Theta} \sum_{i=1}^{n} \log(f(X_i \mid \theta)).$$

Intuitively, the value of θ that maximizes the likelihood (or log likelihood) should be a good estimator of θ_0, because presumably the values of $(X_1, ..., X_n)$ were picked from an area of high probability.

Definition 12.14 (Fisher information)**.** The Fisher information is

$$\mathcal{I}(\theta) = E_X\left(\left(\frac{\partial}{\partial\theta}\log(f(X \mid \theta))\right)^2\right) = E_X\left(-\frac{\partial^2}{\partial\theta^2}\log(f(X \mid \theta))\right).$$

The second equality follows from regularity conditions.

Theorem 12.15. *Under regularity conditions, the Fisher information determines the (approximate) variance of the MLE, which gives an idea of the MLE's accuracy.*

Half-formally, if $X_1, ..., X_n \sim_{\text{i.i.d.}} f(\cdot \mid \theta_0)$, and $\hat{\theta}$ is the MLE, then under regularity conditions,

$$(\text{Distr. of } \hat{\theta}) \approx N\left(\theta_0, \frac{1}{n\mathcal{I}(\theta_0)}\right).$$

Full-formally, under regularity conditions,

$$\left(\text{Distr. of } \sqrt{n\mathcal{I}(\theta_0)}\cdot(\hat{\theta}-\theta_0)\right) \to N(0,1) \text{ as } n\to\infty.$$

This means that for any $x \in \mathbb{R}$,

$$\lim_{n\to\infty} P\left(\sqrt{n\mathcal{I}(\theta_0)}\cdot(\hat{\theta}-\theta_0) \le x\right) = \Phi(x).$$

If θ_0 is unknown, then the same statements hold with $\hat{\theta}$ in place of θ_0, i.e.

$$\lim_{n\to\infty}\left(\text{Distr. of } \sqrt{n\mathcal{I}(\hat{\theta})}\cdot(\hat{\theta}-\theta_0)\right) = N(0,1), \text{ and}$$
$$\lim_{n\to\infty} P\left(\sqrt{n\mathcal{I}(\hat{\theta})}\cdot(\hat{\theta}-\theta_0) \le x\right) = \Phi(x).$$

By this theorem, high Fisher information implies lower variance, and hence better "accuracy" of the estimator.

The theorem above allows us to create confidence intervals for the true parameter using the MLE. In particular, given the critical $z_{\alpha/2} = \Phi^{-1}(1-\frac{\alpha}{2})$, the asymptotic normality of the MLE gives that

$$P\left(\left|\sqrt{n\mathcal{I}(\hat{\theta})}\cdot(\hat{\theta}-\theta_0)\right| > z_{\alpha/2}\right) \approx \alpha, \text{ so } P\left(\theta_0 \in \hat{\theta} \pm z_{\alpha/2}\cdot\frac{1}{\sqrt{n\mathcal{I}(\hat{\theta})}}\right) \approx 1-\alpha.$$

In short, this gives the $1-\alpha$ confidence interval for θ_0 based on the MLE estimator and its asymptotic normality property.

Question 12.16. *How well does the performance of the MLE stack up against other estimators?*

The concept of 'well' or 'goodness' can be approached from the perspective of the MSE, which is comprised of bias and variance. From the above, the MLE is

1. unbiased (asymptotically), and
2. variance approaches $\frac{1}{n\mathcal{I}(\theta_0)}$ (asymptotically).

But if we "fix" one of the two components of the MSE, we arrive at this following lemma.

Lemma 12.17 (Cramer-Rao Inequality)**.** *If $\hat{\theta}$ is an unbiased estimator i.e. $E(\hat{\theta}) = \theta$ for any θ, then*

$$\text{Var}(\hat{\theta}) \geq \frac{1}{n\mathcal{I}(\theta_0)}.$$

*This means that given an unbiased estimator, the variance of the MLE (asymptotically) is **at least** as good as the variance of the unbiased estimator.*

Remark 12.18. However, the Cramer-Rao inequality does not imply that the MLE is the optimal estimator; there may be cases where we are willing to trade bias for variance.

12.2 Bayesian Inference

Question 12.19. *How can you come up with estimators in the Bayesian framework?*

Recall the hierarchical model:

$$\begin{cases} \theta \sim g(\theta) \\ X_1, ..., X_n \mid \theta \sim_{\text{i.i.d.}} f(\cdot \mid \theta) \end{cases}.$$

Then the posterior PMF/density is

$$\begin{aligned} h(\theta \mid X_1, ..., X_n) &= \frac{f(X_1, ..., X_n \mid \theta) \cdot g(\theta)}{\text{marginal distr. of } X_1, ..., X_n} \\ &= (\text{term that doesn't depend on } \theta) \cdot g(\theta) \cdot \prod_{i=1}^{n} f(X_i \mid \theta). \end{aligned}$$

Knowing the posterior PMF/density allows us to calculate potential estimators and credible intervals.

12.2.1 Potential estimators & Accuracy

1. Posterior mean i.e. $\hat{\theta} = E(\theta \mid X_1, ..., X_n)$.
2. Posterior mode i.e. $\hat{\theta} = \text{argmax}_{\theta \in \Theta}\ h(\theta \mid X_1, ..., X_n)$.

Question 12.20. *How do we measure performance of estimator in Bayesian framework?*

The answer isn't as simple as using the MSE, because the expected value of the MSE does not take into account the randomness of the parameter; in other words, the MSE is a performance metric for an estimator into the frequentist framework.

Definition 12.21 (Bayes' risk)**.** The performance metric of an estimator in the Bayesian framework.

Definition 12.22 (Bayes' rule)**.** The estimator $\hat{\theta}$ that minimizes the Bayes' risk.

There exist different performance metrics in the Bayesian framework, so we highlight pertinent ones below.

1. The squared loss Bayes' risk is

$$E((\hat{\theta}-\theta)^2) = E_{\text{marg. distr. of } X_1,...,X_n}(E_{\theta|X_1,...,X_n}((\hat{\theta}-\theta)^2 \mid X_1,...,X_n)).$$

2. The absolute loss Bayes' risk is

$$E(|\hat{\theta}-\theta|) = E_{\text{marg. distr. of } X_1,...,X_n}(E_{\theta|X_1,...,X_n}(|\hat{\theta}-\theta| \mid X_1,...,X_n)).$$

3. The 0/1 loss Bayes' risk is used in the discrete setting, and it measures performance by asking how likely it is that the estimator will not equal the true parameter value.

$$E(\mathbb{1}_{\hat{\theta}\neq\theta_0}) = P(\hat{\theta} \neq \theta).$$

Theorem 12.23. *For each of the Bayes' risks defined above, the according Bayes' rules are*

1. *Squared loss → posterior mean.*
2. *Absolute loss → posterior median.*
3. *0/1 loss → posterior mode.*

Proof. We show only the squared loss case. Observe that

$$\begin{aligned} E((T-t)^2) &= \text{Var}(T-t) + E(T-t)^2 \\ &= \text{Var}(T) + (E(T)-t)^2. \end{aligned}$$

Hence $E((T-t)^2)$ is minimized by setting $t = E(T)$. In the same way, setting $\hat{\theta} = E(\theta \mid X_1,...,X_n)$ minimizes the squared loss. □

12.2.2 Credible interval

Recall the definition of a credible interval: The $1-\alpha$ credible interval I is the interval such that

$$P(\theta \in I \mid X_1,...,X_n) = 1-\alpha.$$

In other words, I is the interval such that the probability that the parameter lies in I given the data $X_1,...,X_n$ is $1-\alpha$.

There exist two methods of constructing a $1-\alpha$ credible interval: the equal-tailed interval & the high-posterior density interval. The two are equivalent if the posterior density is symmetric and single-peaked (i.e. unimodal).

1. **(Equal tailed interval)** Let F_{post} be the CDF of the posterior distribution, and assume F_{post} is continuous. Then the bounds of the $1-\alpha$ equal tailed credible interval I are given by

$$F_{\text{post}}^{-1}\left(\frac{\alpha}{2}\right) \leq \theta \leq F_{\text{post}}^{-1}\left(1-\frac{\alpha}{2}\right).$$

2. **(High-posterior density interval)** Let I be the interval such that

$$\{\theta : f(\theta \mid X_1,...,X_n) \geq c\}, \text{ where } c \text{ is chosen such that } P(I) = 1-\alpha.$$

13 Hypothesis Testing

Given data from a parametric model i.e. data from $f(\cdot \mid \theta)$ with unknown θ, we set up a **binary** model, where we use data to choose between two statements on θ.

Definition 13.1. Let H_0, H_1 be two statements on θ. H_0 is called the null hypothesis, and H_1 is called the alternative hypothesis. If a hypothesis proposes specific values for θ, then it is simple. If a hypothesis proposes a range of values for θ, then it is composite.

For example, $H_0 = 1$ is simple, while $H_1 \neq 1$ is composite. $H_0 = 1$ is simple, and $H_1 = 2$ is simple also.

Definition 13.2 (Hypothesis Test)**.** A hypothesis test is a function that maps from the observed values to $\{H_0, H_1\}$.

Definition 13.3 (Type I, Type II, Power errors)**.**
The notation is in parentheses. The language of "selecting" refers to the output of the hypothesis test.

1. Type I error(α) = probability of selecting H_1 given that H_0 is true.
2. Type II error(β) = probability of selecting H_0 given that H_1 is true.
3. Power = probability of selecting H_1 given that H_1 is true $= 1 - \beta$.

Definition 13.4 (Test statistic)**.** The test statistic is a function from the observed data to a single value such that a decision can be made (i.e. choosing H_0 or H_1).

Definition 13.5 (Null distribution)**.** The distribution of the test statistic, assuming that the null hypothesis is true.

Definition 13.6 (Rejection region)**.** The rejection region consists of the values of the test statistic for which we reject the null hypothesis.

13.1 Designing hypothesis tests

In practice, the general recipe for designing hypothesis tests is as follows:

1. Choose a test statistic T and an error rate α.
2. Set a rejection region R such that $P_{H_0}(T \in R) = \alpha$.
3. If applicable, compute $P_{H_1}(T \notin R) = \beta$.

13.1.1 Conventions for choosing null/alternative hypothesis

- If one hypothesis is simple and the other is composite, set the null to be the simple hypothesis.
- If the result we want to prove is likely true, then set it as the alternative hypothesis. This is because our goal in hypothesis testing is to disprove/reject the null.

13.1.2 Common types of rejection regions

1. One-sided rejection regions : $(c, \infty), [c, \infty), (-\infty, c), (-\infty, c]$.
2. Two-sided rejection regions : $(-\infty, c_1) \cup (c_2, \infty), (-\infty, c_1] \cup [c_2, \infty)$.

Example 13.7. Suppose we have $X \sim$ Exponential(λ). We want to test $H_0 : \lambda = 20$ and $H_1 : \lambda < 20$. Following the steps of the the general recipe we reason:

1. Set X itself as the test statistic, and let $\alpha = 0.1$.

2. Since the probability of getting a certain value generally increases as λ decreases, choose a one-sided rejection region $(c, +\infty)$ such that

$$P(X \in (c, +\infty) \mid \lambda = 20) = P(X > c \mid \lambda = 20) = 0.1.$$

Using the inverse CDF of the exponential distribution, we get $R = (0.1151, \infty)$. Hence if $X \leq 0.1151$, choose H_0; otherwise, choose H_1. In sum, we get the following hypothesis test:

$$\begin{cases} H_0 & \text{if } X \leq 0.1151 \\ H_1 & \text{otherwise.} \end{cases}$$

3. Choose arbitrary values for λ for H_1. Then

- $(\lambda = 10)$ $P(X \leq 0.1151 \mid \lambda = 10) = 0.684 \implies$ power $= 0.316$.
- $(\lambda = 1)$ $P(X \leq 0.1151 \mid \lambda = 1) = 0.109 \implies$ power $= 0.891$.

Intuitively, the Type II error decreases because we specifically chose the rejection region R on the basis that the null is true, so the further the true λ is away from 20, the test statistic X is more likely to fall inside the rejection region.

13.1.3 Likelihood Ratio Test (LRT)

The likelihood ratio test is a way to design a hypothesis test when we have two simple hypotheses. Assume the data comes from a parametric family $f(\cdot \mid \theta)$, and we are testing $H_0 : \theta = \theta_1$ and $H_1 : \theta = \theta_2$. Compute the ratio of the likelihoods of each θ value

$$LR = \frac{\text{Likelihood of } \theta_1}{\text{Likelihood of } \theta_2}.$$

The likelihood may be just $f(x \mid \theta)$, or $\prod f(x_i \mid \theta)$, depending on the data. Design the hypothesis test as follows:

$$\begin{cases} H_0 & \text{if LR} > c \\ H_1 & \text{if LR} \leq c. \end{cases}$$

Note that the cases could switch, depending on the situation.

This next lemma shows that the LRT is the best possible hypothesis test for testing two simple hypotheses.

Lemma 13.8 (Neyman-Pearson Lemma)**.** *Suppose H_0, H_1 are simple hypotheses, and let $c \geq 0$. Suppose α, β be the Type I, Type II errors (respectively) for the LRT with threshold c. Then for any other hypothesis test, if its Type I error $= \alpha$, then its Type II error $\geq \beta$.*

Similarly, for any other hypothesis test, if its Type II error $= \beta$, then its Type I error $\geq \alpha$.

13.1.4 Generalized Likelihood Test

As the name suggests, the generalized likelihood test is for testing two hypotheses that are not necessarily simple. Assume the data comes from a parametric family $f(\cdot \mid \theta)$, and we are testing $H_0 : \theta \in \Omega_0$ and $H_1 : \theta \in \Omega_1$.

The generalized likelihood ratio (GLR) is

$$\Lambda = \frac{\max_{\theta \in \Omega_0} \prod_{i=1}^n f(X_i \mid \theta)}{\max_{\theta \in \Omega_0 \cup \Omega_1} \prod_{i=1}^n f(X_i \mid \theta)}.$$

To find the GLR, first calculate the MLE (since the MLE is the value that gives the maximum) given the appropriate space i.e. Ω_0, Ω_1, then plug the MLE back in.

Proposition 13.9.

1. $\Lambda \leq 1$.
2. *If $\Lambda \approx 1$, then we do not reject the null.*
3. *If Λ is far away from 1 by a certain threshold amount, then we reject the null.*

Proof.

1. If $A \subset B$, then $\sup A \leq \sup B$, and the result follows.
2. If the null is actually true, then the maximum will be achieved from inside Ω_0, which means that the numerator and the denominator should be around the same value.
3. If instead the alternative is true, then the denominator should be significantly greater than the numerator.

□

To decide the threshold, we use an asymptotic result involving Λ.

Proposition 13.10. *Let Λ be the GLR. Under some regularity conditions, $-2\log(\Lambda)$ asymptotically follows a $\chi^2_{d-d_0}$ distribution, where d, d_0 are the dimensions of $\Omega_0 \cup \Omega_1, \Omega_0$ respectively.*

In particular, the regularity condition is that Ω_0 must lie in the interior of $\Omega_0 \cup \Omega_1$ (in the topological sense, so there must exist an ϵ-thick wall around Ω_0 such that it still fits inside $\Omega_0 \cup \Omega_1$).

To run a generalized likelihood ratio test at level α, we need to compute a threshold such that

$$P_{H_0}(\Lambda < (\text{threshold})) \approx \alpha.$$

Hence, by the proposition, we can simply set

$$\text{threshold} = F^{-1}_{\chi^2_{d-d_0}}(1-\alpha), \text{ so } \begin{cases} H_0 & -2\log(\Lambda) \leq \text{threshold} \\ H_1 & -2\log(\Lambda) > \text{threshold}. \end{cases}$$

Similarly, for a calculated Λ, its p-value is $1 - F_{\chi^2_{d-d_0}}(-2\log(\Lambda))$.

13.2 p-values

Definition 13.11 (p-value)**.** Let T be a test statistic, and R be the rejection region of a hypothesis test. Then for any value t of T, the p-value of t is the value of α for which t lies on the boundary of R.

In other words, the p-value of $T = t$ is the α level at which t switches from being outside the rejection region to being inside it (or vice versa).

Another interpretation of the p-value of $T = t$ is the probability that the test statistic is at least as extreme as t, given that H_0 is true.

Example 13.12. Suppose we have $X \sim$ Exponential(λ). Testing $H_0 : \lambda = 20$ versus $H_1 : \lambda < 20$, we concluded that a one-sided rejection region (c, ∞) would be most appropriate. What is the p-value of x?

Since X itself is the test statistic, we solve

$$P(X > x \mid \lambda = 20) = e^{-20x} = \alpha.$$

This equation aligns with both interpretations of p-value in the definition. Hence, after observing X, the p-value is e^{-20X}.

14 Confidence Intervals, Hypothesis Testing, & p-values

Theorem 14.1. *For a two-sided test / confidence interval, the following are equivalent:*

1. *If we test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, then the test statistic falls into a rejection region R with level α.*
2. *If we test $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, then the p-value of the test statistic is $\leq \alpha$.*
3. *If we construct a $1 - \alpha$ confidence interval for θ, then θ_0 falls outside of the confidence interval.*

The same is true for one-sided tests and one-sided confidence intervals

Proof. We present only the intuitive ideas.

- ((1) $\iff$ (2)) Notice that as the value of the test-statistic increases, its p-value decreases. This means that if the test-statistic is so large that it falls inside R, then it will have a lower p-value than α.
- ((1) $\iff$ (3)) Suppose the $1 - \alpha$ confidence interval gives

$$P(\text{bound}_1(\theta) \leq \theta_0 \leq \text{bound}_2(\theta)) = 1 - \alpha.$$

  We can rewrite these bounds to get

$$P(\text{bound}_1(\theta_0) \leq \theta \leq \text{bound}_2(\theta_0)) = 1 - \alpha.$$

  Hence the probability that θ falls outside of the bounds, given that the null is true, is α, which means that the test statistic falls inside R with level α.

□

15 Multinomial Data

Definition 15.1. The multinomial distribution is a generalization of the binomial distribution, where

- we have $m \geq 2$ categories (binomial distribution is $m = 2$),
- the probabilities $p_1, ..., p_m$ sum to 1 ($p_1 + ... + p_m = 1$), and
- we draw n independent observations, which each obey $p_1, ..., p_m$, and we count X_i for $i = 1, ..., m$, where X_i is the number of observations that fall into the ith category ($X_1 + ...X_m = n$).

Definition 15.2 (Probability simplex)**.** For a multinomial distribution with m categories, we define the probability simplex

$$\Delta_m = \{(p_1, ..., p_m) \mid p_i \in [0, 1] \text{ s.t. } p_1 + ... + p_m = 1\} \subset \mathbb{R}^m.$$

We want to run hypothesis tests of the for $H_0 : (p_1, ..., p_m) \in \Omega_0$ vs. $H_1 : (p_1, ..., p_m \in \Delta_m \setminus \Omega_0$. Since H_1 is not a simple hypothesis, we can use either the generalized likelihood ratio test, or Pearson's χ^2 test.

1. **Generalized Likelihood Ratio Test.** The likelihood of the multinomial distribution, after having observed $X_1, ..., X_m$ (recall that $X_1, ..., X_n$ are not i.i.d. random variables, but rather the observed counts on each of the categories), and parameters $p_1, ..., p_m$, is

$$\frac{n!}{\prod_{i=1}^m X_i} \prod_{i=1}^m p_i^{X_i}.$$

   To find the GLR, we have the general strategy for finding the MLE within a space Ω_0:

(a) Find the dimension i.e. the number of free parameters of Ω_0.

(b) Rewrite $(p_1, ..., p_m)$ as a function of the free parameters.

(c) Rewrite the log likelihood as a function of the free parameters.

(d) Solve for each of the free parameters by taking the partial derivative w.r.t. each free parameter, and setting to 0. This gives you the MLE for each parameter.

(e) Plug the MLE back into the likelihood function.

In general though, the MLE of the likelihood within the entire probability simplex is

$$\hat{p_1} = \frac{X_1}{n}, \hat{p_2} = \frac{X_2}{n}, ..., \hat{p_m} = \frac{X_m}{n}.$$

Having computed the GLR, in order to test H_0 or the p-value, compare $-2\log(\Lambda)$ against $\chi^2_{d-d_0}$.

2. **Pearson's χ^2 Test.** To generate the test statistic for the Pearson χ^2 test, we have

(a) Calculate the expected counts for each category, i.e. for $i = 1, ..., m$,

$$E_i = n \cdot \hat{p_i}, \text{ where } \hat{p_i} \text{ is the MLE.}$$

(b) Calculate the test statistic

$$X^2 = \sum_{i=1}^{n} \frac{(X_i - E_i)^2}{E_i}.$$

The division by E_i is there because the size of a discrepancy is relative to the number of samples taken. For example, if $n = 10$, then the discrepancy between 4 and 7 is significant, but if $n = 10000$, then the discrepancy between 100 and 103 is not that significant.

(c) X^2 is asymptotically distributed to $\chi^2_{d-d_0}$, so to test H_0 and calculate p-values, compare X^2 against $\chi^2_{d-d_0}$.

A Cheat Sheet

A.1 Discrete distribution (PMF — CDF — EV — Var)

1. Bernoulli(p): $\emptyset \mid \emptyset \mid p \mid p(1-p)$
2. Binomial(n, p): $\binom{n}{k}p^k(1-p)^{n-k} \mid \emptyset \mid np \mid np(1-p)$
3. Geometric(p): $(1-p)^k p \mid 1-(1-p)^k \mid \frac{1}{p} \mid \frac{1-p}{p^2}$.
4. Poisson(λ): $\frac{\lambda e^{-\lambda}}{k!} \mid \emptyset \mid \lambda \mid \lambda$

A.2 Continuous distribution (Dens — CDF — EV — Var)

1. Uniform$[a, b]$: $\frac{1}{b-a} \mid \frac{x-a}{b-a} \mid \frac{a+b}{2} \mid \frac{1}{12}(b-a)^2$
2. Exponential(λ): $\lambda e^{-\lambda x} \mid 1-e^{-\lambda x} \mid \frac{1}{\lambda} \mid \frac{1}{\lambda^2}$
3. Normal(μ, σ^2): $\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2} \mid \emptyset \mid \mu \mid \sigma^2$

A.3 Expectation (μ)

1. (Discrete) $E(X) = \sum_x x \cdot p_X(x)$,
2. (Continuous) $E(X) = \int_{-\infty}^{\infty} x \cdot f_X(x)\ dx$
3. (Function) $E(g(X)) = \sum_x g(x) \cdot p_X(x)$ or $\int_{-\infty}^{\infty} g(x) f_X(x)\ dx$
4. $E(\mathbb{1}_A) = P(A)$.
5. $E(a + b_1X_1 + ... + b_nX_n) = a + b_1E(X_1) + ... + b_nE(X_n)$.
6. $X \leq Y$ almost surely $\implies E(X) \leq E(Y)$.
7. $X \in [0, \infty), \forall t > 0 \implies P(X \geq t) \leq \frac{E(X)}{t}$.
8. Joint $(X, Y) \implies E(X) = E_Y(E_{X|Y}(X \mid Y))$
9. $X \perp Y \implies E(XY) = E(X)E(Y)$
10. $X \perp Y \implies g(X) \perp h(Y) \implies E(g(X)h(Y)) = E(g(X)) \cdot E(h(Y))$.

A.4 Variance (σ^2)

1. $\text{Var}(X) = E((X - \mu_X)^2)$.
2. (Discrete) $\text{Var}(X) = \sum_x (x - \mu_X)^2 p_X(x)$.
3. (Continuous) $\text{Var}(X) = \int_{-\infty}^{\infty} (x - \mu_X)^2 f_X(x) dx$.
4. $\text{Var}(X) = E(X^2) - E(X)^2$
5. RV X and $\forall t > 0 \implies P(|X - \mu_X| \geq t) \leq \frac{\sigma_X^2}{t^2}$.
6. $\text{Var}(X) = 0$ if and only if $P(X = \mu_X) = 1$.
7. $\text{Var}(a + bX) = b^2\text{Var}(X)$.
8. $X_1, ..., X_n$ mutually independent $\implies \text{Var}(X_1 + ... + X_n) = \text{Var}(X_1) + ... + \text{Var}(X_n)$.

A.5 Covariance

1. $\text{Cov}(X,Y) = E((X-\mu_X)\cdot(Y-\mu_Y))$.
2. $\text{Cov}(X,Y) = E(XY) - E(X)E(Y)$.
3. $\text{Corr}(X,Y) = \frac{\text{Cov}(X,Y)}{\sigma_X \sigma_Y}$.
4. $\text{Cov}(X,X) = \text{Var}(X)$, and $\text{Corr}(X,X) = 1$.
5. $\forall X, Y, \text{Corr}(X,Y) \in [-1,1]$.
6. $\text{Corr}(X,Y) = \pm 1 \iff Y = aX + b$
7. $\text{Cov}(a+bX, a'+b'Y) = bb'\text{Cov}(X,Y)$
8. $b, b' \neq 0 \implies \text{Corr}(a+bX, a'+b'Y) = \text{Corr}(X,Y) \cdot \text{ sign}(bb')$
9. $\text{Cov}(X_1 + ... + X_n, Y_1 + ... + Y_n) = \sum_{i=1}^{n} \sum_{j=1}^{m} \text{Cov}(X_i, Y_j)$.
10. $X \perp Y \implies \text{Cov}(X,Y) = \text{Corr}(X,Y) = 0$.

A.6 Tricks

1. Necessary condition for independent RV is that the supports match.
2. Use tower law in a joint distribution setting.
3. Use tower law to remove ambiguity and set conditions on RV.
4. The idea of taking one condition and iterating over all possible conditions is a useful one. (basis for law of total probability and tower law).
5. For continuous, derive CDF first, then differentiate to get the density function.
6. Conditional distributions are not so different from the ordinary distributions. Consider $X \mid Y = y$, just observe all the values of (X,Y) and keep the $X's$ that have $Y = y$ associated to them.
7. Remember whether RV is discrete or continuous.
8. Remember which variables are constants (especially useful for anything related to expectation.)
9. Remember what the support is.
10. If $\{B_1, ..., B_n\}$ partition the event A, then

$$P(B_1 \mid A) + ... + P(B_n \mid A) = 1.$$

In particular, this means $P(B \mid A) = 1 - P(B^c \mid A)$, so $P(B \mid A)P(A) = P(A) - P(B^c \mid A)P(A)$.

A.7 Analogs

1. (Summing over every condition / conditioning, then summing)

$$\longleftrightarrow P(A) = \sum_i P(A \mid B_i)P(B_i)$$

$$\longleftrightarrow P(A) = E_X(P(A|X)) = \int_{x=-\infty}^{\infty} P(A \mid X) f_X(x) dx$$

$$\longleftrightarrow f_X(x) = \int_{y=-\infty}^{\infty} f_{X,Y}(x,y) dy = \int_{y=-\infty}^{\infty} f_{X|Y}(x \mid y) f_Y(y)\ dx$$

$$\longleftrightarrow p_X(x) = \sum_y p_{X,Y}(x,y) = \sum_y p_{X|Y}(x \mid y) p_Y(y)$$