STAT 24500 Cheatsheet

Hye Woong Jeon

Contents

1 Poisson Distribution

The Poisson distribution is the distribution used for occurrences of **rare** events.

Definition 1.1 (Poisson Distribution)**.** Assuming $X \sim \text{Poisson}(\lambda)$, we have

- **PMF:** $e^{-\lambda}\frac{\lambda^k}{k!}$ where $k = 0, 1, 2, ...$

- **Expected value:** λ.
- **Variance:** λ.

Check. We check that each of components in the definition above are true.

- **PMF.** To check that the PMF is indeed a probability distribution, we need to check that all the probabilities sum to 1. Hence,

$$\sum_{k=0}^{\infty} e^{-\lambda}\frac{\lambda^k}{k!} = e^{-\lambda}\sum_{k=0}^{-\infty}\frac{\lambda^k}{k!} = e^{-\lambda}\cdot e^{\lambda} = \boxed{1}.$$

- **Expected value.** We have

$$E(X) = \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda e^{-\lambda}\sum_{k=0}^{\infty}\frac{\lambda^{k-1}}{(k-1)!} = \lambda e^{-\lambda}\sum_{k=1}^{\infty}\frac{\lambda^k}{(k)!} = \boxed{\lambda}.$$

- **Variance.** Since $\mathrm{Var}(X) = E(X^2) - E(X)^2$, we need to find $E(X^2)$. We have

$$E(X^2) = \sum_{k=0}^{\infty} k^2e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=0}^{\infty}(k^2-k+k)e^{-\lambda}\frac{\lambda^k}{k!} = \sum_{k=0}^{\infty}(k(k-1))e^{-\lambda}\frac{\lambda^k}{k!} + \sum_{k=0}^{\infty} ke^{-\lambda}\frac{\lambda^k}{k!} = \lambda^2+\lambda.$$

  Hence $\mathrm{Var}(X) = E(X^2) - E(X)^2 = \boxed{\lambda}$.

□

Concept 1.2. *What constitutes a rare event? A rare event can be characterized by a Bernoulli random variable with sufficiently small p. But how small?*

We can call an event rare if its probability of success p is such that $\lim_{n\to\infty} np < \infty$. *Intuitively, this means that p is so small that it decays faster than n increases.*

Theorem 1.3 (Law of Small Numbers)**.** *If* $np \to \lambda$ *as* $n \to \infty$, *then*

$$\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k} = e^{-\lambda}\frac{\lambda^k}{k!}.$$

In other words, the binomial distribution of an exceedingly rare event will converge to the Poisson distribution.

Proof. We make three observations:

1. $\lim_{n\to\infty}\dfrac{n!}{(n-k)!\,n^k} = \lim_{n\to\infty}\left(\dfrac{n}{n}\right)\cdot\left(\dfrac{n-1}{n}\right)\cdot\ldots\cdot\left(\dfrac{n-(k-1)}{n}\right) = 1.$
2. $\lim_{n\to\infty}(np)^k = \lambda^k.$
3. $\lim_{n\to\infty}\left(1-\dfrac{np}{n}\right)^{n-k} = \lim_{n\to\infty}\left(1-\dfrac{np}{n}\right)^{n\cdot\frac{n-k}{n}} \approx \left(1-\dfrac{\lambda}{n}\right)^{n} \to e^{-\lambda}.$

From these three observations, we have

$$\begin{aligned}\lim_{n\to\infty}\binom{n}{k}p^k(1-p)^{n-k} &= \lim_{n\to\infty}\frac{n!}{(n-k)!\cdot k!}p^k(1-p)^{n-k}\\ &= \lim_{n\to\infty}\frac{1}{k!}\cdot\frac{n!}{(n-k)!\cdot n^k}\cdot(np)^k\cdot\left(1-\frac{np}{n}\right)^{n-k}\\ &= \boxed{e^{-\lambda}\frac{\lambda^k}{k!}}.\end{aligned}$$

□

Proposition 1.4. *If $X_1 \sim \text{Poisson}(\lambda_1)$ and $X_2 \sim \text{Poisson}(\lambda_2)$ are independent, then*

$$X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2).$$

Proof. As a sanity check, we confirm that $E(X_1 + X_2) = \text{Var}(X_1 + X_2) = \lambda_1 + \lambda_2$. For rigor, we have

$$\begin{aligned}\mathbb{P}(X_1 + X_2 = k) &= \sum_{l=0}^{k}\mathbb{P}(X_1 = l, X_2 = k-l)\\ &= \sum_{l=0}^{k}\mathbb{P}(X_1 = l)\cdot\mathbb{P}(X_2 = k-l)\\ &= e^{-(\lambda_1+\lambda_2)}\sum_{l=0}^{k}\frac{\lambda_1^k}{l!}\cdot\frac{\lambda_2^{k-l}}{(k-l)!}\\ &= e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!}\sum_{l=0}^{k}\underbrace{\frac{k!}{(\lambda_1+\lambda_2)^k}\cdot\frac{\lambda_1^k}{l!}\cdot\frac{\lambda_2^{k-l}}{(k-l)!}}_{\text{PMF of Binomial}(k,\frac{\lambda_1}{\lambda_1+\lambda_2})} = \boxed{e^{-(\lambda_1+\lambda_2)}\frac{(\lambda_1+\lambda_2)^k}{k!}}.\end{aligned}$$

Hence $X_1 + X_2 \sim \text{Poisson}(\lambda_1 + \lambda_2)$. □

Proposition 1.5. *For $X_1, ..., X_n \sim_{\text{i.i.d.}} \text{Poisson}(\lambda)$ for some λ, the MLE for λ is $\bar{X}$.*

Proof. Use the typical MLE method: joint likelihood → log the likelihood → differentiate and set to 0 → solve for λ. □

2 Methods of CI Construction

Recall that a $(1-\alpha)\%$ confidence interval is an interval $[\lambda_{\text{left}}, \lambda_{\text{right}}]$ that has a $1-\alpha$ probability of covering the true parameter λ.

For this section, we work with confidence intervals for estimators of the Poisson distribution. Hence, recall that the MLE of the Poisson distribution is the sample mean.

2.1 Wald's Method

Theorem 2.1. *Using the CLT, the law of large numbers, and Slutsky's theorem, the $(1-\alpha)\%$ confidence interval is*

$$\left[\hat{\lambda} - z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{\lambda}}{n}}, \hat{\lambda} + z_{1-\frac{\alpha}{2}}\cdot\sqrt{\frac{\hat{\lambda}}{n}}\right], \text{ where } z_{1-\frac{\alpha}{2}} = \Phi^{-1}\left(1-\frac{\alpha}{2}\right).$$

Proof. By the CLT, $\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}} \to N(0,1)$. By the law of large numbers, $\hat{\lambda} \to \lambda$. Hence Slutsky's theorem gives that

$$\frac{\hat{\lambda}-\lambda}{\sqrt{\hat{\lambda}/n}} \to N(0,1).$$

From this follows the $(1-\alpha)\%$ confidence interval shown above. To note some observations, observe that

1. As $n \to \infty$, the CI gets narrower and hence more accurate.
2. As $\alpha \to 0$, the CI gets wider and its length goes to ∞.

□

While being the simplest method, Wald's method is unfortunately not very accurate, given its dependence on both the CLT and the law of large numbers.

2.2 Wilson's Method

Wilson's method depends solely on the CLT, and hence removes one of the sources of estimation from Wald's method. This makes Wilson's method more accurate that Wald's.

Theorem 2.2. *Using the CLT, the $(1-\alpha)\%$ confidence interval is given by the solutions (solve for λ) to the equation*

$$(\hat{\lambda}-\lambda)^2 = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}.$$

Proof. By the CLT, $\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}} \to N(0,1)$. Hence the event

$$\begin{aligned}\mathbb{P}\left(\left|\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}}\right| < z_{1-\frac{\alpha}{2}}\right) &= \mathbb{P}\left(\frac{(\hat{\lambda}-\lambda)^2}{\lambda/n} < z_{1-\frac{\alpha}{2}}^2\right) \\ &= \mathbb{P}\left((\hat{\lambda}-\lambda)^2 < z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}\right) \approx 1-\alpha.\end{aligned}$$

Now, $(\hat{\lambda}-\lambda)^2 < z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}$ if and only if λ is in between the solutions for the equation $(\hat{\lambda}-\lambda)^2 = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}$, solved for λ. Hence if $f(\lambda) = (\hat{\lambda}-\lambda)^2$, then the $(1-\alpha)\%$ confidence interval is

$$\left[\text{sol}_1\left(f(\lambda) = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}\right), \text{sol}_2\left(f(\lambda) = z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}\right)\right].$$

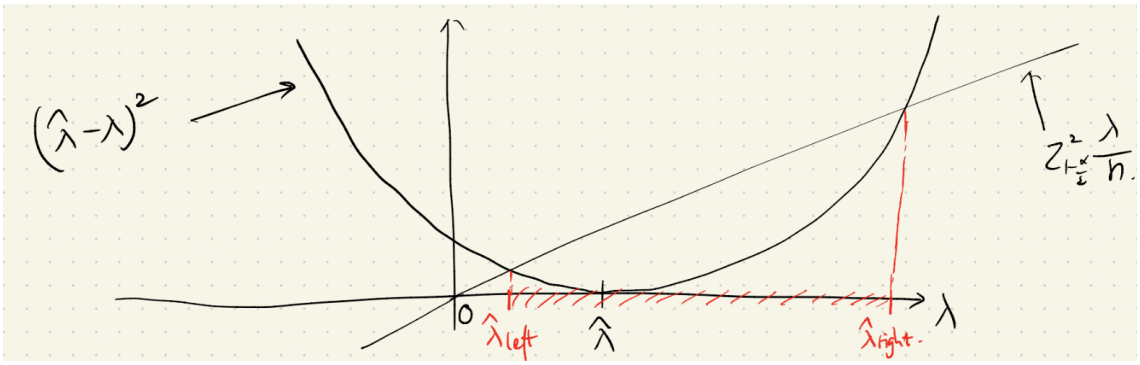
Notice that as $n \to \infty$, the slope of $z_{1-\frac{\alpha}{2}}^2 \cdot \frac{\lambda}{n}$ decreases, hence making the CI narrower. Furthermore, this CI is asymmetrical and for good reason; the parameter for the Poisson distribution is strictly positive, so it makes sense for the CI to be longer on the right. □

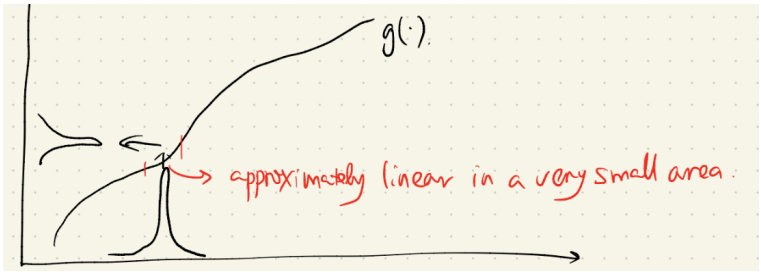
2.3 Variance Stability Transformation (VST)

Proposition 2.3. *Suppose g is a differentiable function. Then if X converges to a normal distribution with small variance, then $g(x)$ converges to a normal distribution also.*

Proof. We provide only an intuitive proof. Because g is differentiable, it is approximately linear within the small variance of X. Since the linear transformation of a normally distributed random variable is also normally distributed, $g(X)$ converges to a normal distribution also.

□





The problem with Wald and Wilson's methods is the variance's dependence on the true parameter λ. The VST aims to resolve this by applying a differentiable transformation to $\hat{\lambda}$. More specifically, the setup from the CLT is

$$\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}} \to N(0,1), \text{ so } \sqrt{n}\cdot(\hat{\lambda}-\lambda) \to N(0,\lambda).$$

We can't construct a CI based on this because the convergent normal distribution depends on λ. The VST solves this by supplying a differentiable function g such that

$$\sqrt{n}\cdot(g(\hat{\lambda})-g(\lambda)) \to N(0,1).$$

Theorem 2.4. *The differentiable function g that works for the Poisson distribution is $g(x)=2\sqrt{x}$. From this, the $(1-\alpha)\%$ event (and thus interval) is*

$$\sqrt{\lambda} \in \left[\sqrt{\hat{\lambda}} - \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}, \sqrt{\hat{\lambda}} + \frac{z_{1-\frac{\alpha}{2}}}{2\sqrt{n}}\right].$$

Proof. We want to find a g such that $\sqrt{n}\cdot(g(\hat{\lambda})-g(\lambda)) \to N(0,1)$. By the law of large numbers, since $\hat{\lambda}\to\lambda$, $\hat{\lambda}$ lies in a neighborhood of λ. Therefore, we can apply Taylor's Theorem:

$$g(\hat{\lambda}) \approx g(\lambda) + g'(\lambda)(\hat{\lambda}-\lambda) + \frac{g''(\lambda)}{2}(\hat{\lambda}-\lambda)^2 + \cdots$$

But since $\frac{\hat{\lambda}-\lambda}{\sqrt{\lambda/n}}$ converges to a constant distribution, the top and bottom must have the same order. Therefore, $(\hat{\lambda}-\lambda) \sim \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)$. Hence

$$g(\hat{\lambda}) \approx g(\lambda) + g'(\lambda)(\hat{\lambda}-\lambda) + \mathcal{O}\left(\frac{1}{n}\right).$$

Hence $g(\hat{\lambda}) - g(\lambda) \approx g'(\lambda)(\hat{\lambda} - \lambda)$. Now,

$$\sqrt{n} \cdot (g(\hat{\lambda}) - g(\lambda)) \approx g'(\lambda) \cdot \underbrace{\sqrt{n}(\hat{\lambda} - \lambda)}_{\to N(0,\lambda)} \to N(0, |g'(\lambda)|^2 \lambda).$$

Setting $|g'(\lambda)|^2\lambda = 1$, $g'(\lambda) = \frac{1}{\sqrt{\lambda}} \iff g(\lambda) = 2\sqrt{\lambda}$. Therefore,

$$\sqrt{n} \cdot (2\sqrt{\hat{\lambda}} - 2\sqrt{\lambda}) \to N(0, 1).$$

From this, the appropriate confidence interval can be obtained.

□

3 Exponential Distribution

The exponential distribution has a close relation to the Poisson distribution through the *Poisson process*, which we explore later.

Definition 3.1 (Exponential Distribution)**.** Assuming $X \sim \text{Exponential}(\lambda)$, where $\lambda > 0$, we have

1. **PMF.** $\lambda e^{-\lambda x}$, where $x \in [0, \infty)$.
2. **Expected value.** $\frac{1}{\lambda}$.
3. **Variance.** $\frac{1}{\lambda^2}$.

Proposition 3.2. *For $X_1, ..., X_n \sim_{\text{i.i.d.}} \text{Exponential}(\lambda)$ for some λ, the MLE for λ is $\frac{1}{\bar{X}}$.*

In order to construct confidence intervals for $\hat{\lambda} = \frac{1}{\bar{X}}$, we need to find an asymptotic distribution for $\hat{\lambda}$. Below are two methods:

1. **Delta method.** The MLE $\hat{\lambda}$ is based on the sample mean, and the sample mean converges in distribution to the normal distribution. In other words, we know that $\sqrt{n}(\bar{X} - \frac{1}{\lambda}) \to N\left(0, \frac{1}{\lambda^2}\right)$ by the CLT, and we want $\sqrt{n}(g(\bar{X}) - g\left(\frac{1}{\lambda}\right)) = \sqrt{n}(\hat{\lambda} - g\left(\frac{1}{\lambda}\right)) \to N(0, \text{something})$.

   Notice that $g(t) = \frac{1}{t}$. The delta method gives that

   $$\sqrt{n}(\hat{\lambda} - \lambda) = \sqrt{n}\left(g(\bar{X}) - g\left(\frac{1}{\lambda}\right)\right) \to N\left(0, \left|g'\left(\frac{1}{\lambda}\right)\right|^2 \cdot \frac{1}{\lambda^2}\right) = \boxed{N(0, \lambda^2)}.$$

2. **Fisher information.** Denoting the density function as $p_\theta(x)$, define the *score* as $l_\theta(x) = \frac{\partial}{\partial\theta} p_\theta(x)$. Then the *Fisher information* is $\mathcal{I}_\theta = \mathbb{E}((l_\theta(x))^2)$.

   Recall that $\hat{\lambda} \to N(\lambda, \frac{1}{n\mathcal{I}_\lambda})$. Computing the score and Fisher information of the MLE of exponential, we get that $\sqrt{n}(\hat{\lambda} - \lambda) \to \boxed{N(0, \lambda^2)}$.

3.1 Confidence Intervals for the Exponential Distribution

3.1.1 Wald's Method

Asymptotic convergence gives $\sqrt{n}(\hat{\lambda} - \lambda) \to N(0, \lambda^2)$, the law of strong numbers gives $\hat{\lambda} \to \lambda$. Hence the $(1 - \alpha)$ confidence interval is

$$\lambda \in \left[\hat{\lambda} - z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\lambda}}{\sqrt{n}}, \hat{\lambda} + z_{1-\frac{\alpha}{2}} \cdot \frac{\hat{\lambda}}{\sqrt{n}}\right].$$

3.1.2 Wilson's Method

Asymptotic convergence gives $\sqrt{n}(\hat{\lambda} - \lambda) \to N(0, \lambda^2)$, so

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\hat{\lambda} - \lambda)}{\lambda}\right| < z_{1-\frac{\alpha}{2}}\right) = \mathbb{P}\left(|\hat{\lambda} - \lambda| < z_{1-\frac{\alpha}{2}} \cdot \frac{\lambda}{\sqrt{n}}\right) \approx 1 - \alpha.$$

Note that $\lambda > 0$, so $|\lambda| = \lambda$. Solving the equation $|\hat{\lambda} - \lambda| = z_{1-\frac{\alpha}{2}} \cdot \frac{\lambda}{\sqrt{n}}$ for λ, we get the $(1-\alpha)$ confidence interval

$$\lambda \in \left[\frac{\hat{\lambda}}{1 + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}}, \frac{\hat{\lambda}}{1 - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}}\right].$$

Remark 3.3. How can we be sure that there are two solutions to the equation? The slope on the RHS is $\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}$, and the slope on the LHS is 1. What happens if $\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \geq 1$? Then only the left solution will exist.

However, this is unlikely. In order for $\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}} \geq 1$, $n \leq 3$. If your $n \leq 3$, stop trying to compute CIs and get more data, please.

3.1.3 VST Method

Asymptotic convergence gives $\sqrt{n}(\hat{\lambda} - \lambda) \to N(0, \lambda^2)$, so we want a function g such that $\sqrt{n}(g(\hat{\lambda}) - g(\lambda)) \to N(0, 1)$. Using the delta method, we want to solve $|g'(\lambda)|^2 \cdot \lambda^2 = 1 \implies g'(\lambda) = \frac{1}{\lambda}$. Setting $g = \log$ does the trick. Hence

$$\mathbb{P}(-z_{1-\frac{\alpha}{2}} < \sqrt{n}(\log(\hat{\lambda}) - \log(\lambda))) < z_{1-\frac{\alpha}{2}}) \approx 1 - \alpha.$$

Manipulating the inequality as an expression of $\hat{\lambda}$, we get

$$\lambda \in \left[\hat{\lambda} \cdot e^{-\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}}, \hat{\lambda} \cdot e^{\frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}}\right].$$

4 Poisson Process

There exists a relationship between the Poisson and exponential distributions. First, note the following hierarchy.

Remark 4.1. Schematically speaking, we have the hierarchy

$$\text{random variable} \to \text{random vector} \to \text{random function}.$$

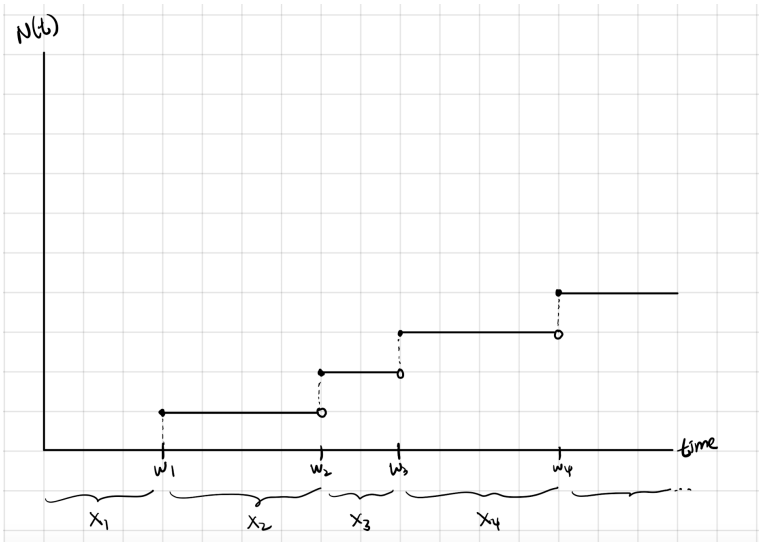
A random vector is made up of multiple random variables, and a random function can be thought of as an infinite-dimensional random vector. (The output of a random function is a random variable.) For example, with the random function $X(t)$, you can build a k-dimensional random vector $(X(t_1), ..., X(t_k))$.

Suppose you are observing the arrival of buses. Denote t as time, and $N(t)$ as the number of buses you have observed until time t. If you plot $N(t)$ against t, you should expect to see a step function, where each step occurs when you see a new bus. This function N can be modeled as a Poisson process.

4.1 Homogeneous Poisson Process

Definition 4.2. Suppose $N(t) \sim PP(\lambda)$, where λ is fixed. Then the following properties hold:

1. $N(0) = 0$.
2. For any $s, t \in \mathbb{R}^+$, $(N(s+t) - N(s)) \perp (N(s))$. In other words, the Poisson process is memoryless; the random variables that come after s are not affected by history.
3. For any $s \in \mathbb{R}^+$, $N(s+t) - N(s) \sim \text{Poisson}(\lambda t)$.



Remark 4.3. We make a few observations, recalling that the Poisson random variable can be seen as the distribution that models rare occurrences.

- By property 3, λ represents the number of expected occurrences in a given time frame. Recall that the expected value of a Poisson random variable is λ, so $\lambda \cdot t$ gives the number of expected occurrences in a time frame of length t.
- The Poisson process can be thought of as the accumulation of Poisson random variables. Since sums of independent Poisson random variables are also Poisson distributed with respect to the sums of the λs, property 3 can be reinterpreted. If $t \in \mathbb{N}$, then

$$\begin{aligned} N(t) &= N(t+0) - N(0) \\ &= \underbrace{N(t+(t-1)) - N(t-1)}_{\sim \text{Poisson}(\lambda)} + \underbrace{N(t-1) - N(t-2)}_{\sim \text{Poisson}(\lambda)} + \ldots + \underbrace{N(1) - N(0)}_{\sim \text{Poisson}(\lambda)} \\ &\sim \text{Poisson}(\lambda t). \end{aligned}$$

Definition 4.4. The $w_1, w_2, \ldots$ in the graph above are called *waiting times.* Notice that $w_1 = X_1, w_2 = X_1 + X_2, \ldots$ In other words the waiting times are the partial sums of $\sum X_i$.

Theorem 4.5. *The times between the waiting times are independently exponentially distributed. In other words,* $X_1, X_2, \ldots \sim_{\text{i.i.d.}} \text{Exponential}(\lambda)$. *Also,* $w_k = X_1 + \ldots + X_k \sim \text{Gamma}(k, \frac{1}{\lambda})$.

Proof. We begin by asking what the distribution of X_1 is. Computing the CDF, we have

$$\mathbb{P}(X_1 \leq t) = 1 - \mathbb{P}(X_1 > t).$$

Notice that $\{X_1 > t\}$ is equivalent to $\{N(t) = 0\}$. Hence we have

$$\mathbb{P}(X_1 \leq t) = 1 - \mathbb{P}(X_1 > t) = 1 - \mathbb{P}(N(t) = 0) = 1 - \mathbb{P}(N(t+0) - N(0) = 0) = 1 - e^{-\lambda t}.$$

This is the CDF of Exponential(λ), so $X_1 \sim$ Exponential(λ). We next ask what the joint distribution of X_1, X_2 is. We first calculate the CDF of the conditional distribution:

$$\mathbb{P}(X_2 \leq t \mid X_1 = s) = 1 - \mathbb{P}(X_2 > t \mid X_1 = s).$$

Notice that $\{X_2 > t \mid X_1 = s\} = \{N(s+t) - N(s) = 0\}$, given that $X_1 = s$. Hence we have

$$\begin{aligned}\mathbb{P}(X_2 \leq t \mid X_1 = s) &= 1 - \mathbb{P}(X_2 > t \mid X_1 = s) \\ &= 1 - \mathbb{P}(N(s+t) - N(s) = 0) = 1 - e^{\lambda t}.\end{aligned}$$

However, notice that the conditional CDF does not depend on s... we will show that the conditional CDF is equivalent to the marginal CDF of X_2:

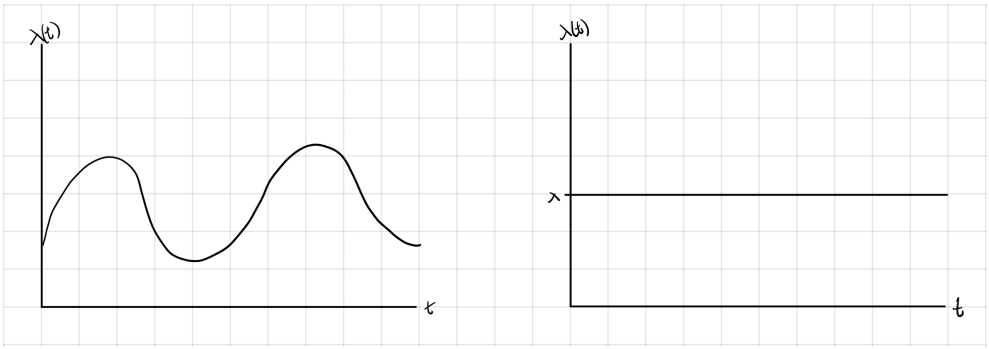
$$\begin{aligned}\mathbb{P}(X_2 \leq t) &= \mathbb{E}(\mathbb{P}(X_2 \leq t \mid X_1 = s)) \\ &= \sum_s \mathbb{P}(X_2 \leq t \mid X_1 = s)\mathbb{P}(X_1 = s) \\ &= \sum_s (1 - e^{\lambda t})\mathbb{P}(X_1 = s) \\ &= (1 - e^{\lambda t}) \sum_s \mathbb{P}(X_1 = s) = 1 - e^{\lambda t} = \mathbb{P}(X_2 \leq t \mid X_1 = s).\end{aligned}$$

This means that the conditional density and marginal densities of X_2 are equivalent, so $X_1, X_2 \sim_{\text{i.i.d.}}$ Exponential(λ).

By induction, it follows that $X_1, X_2, ... \sim_{\text{i.i.d.}}$ Exponential(λ). Since w_k is the sum of identically and independently distributed exponential random variables, $w_k = X_1 + ... + X_k \sim \text{Gamma}(k, \frac{1}{\lambda})$. □

4.2 Inhomogeneous Poisson Process

But what if the rate of occurrences (λ) is not fixed? After all, the rate at which buses arrive during rush hour is different from that at midnight. We can define a rate function $\lambda(t)$ that corresponds to the time, and modify the three properties.



Definition 4.6. Suppose $N(t) \sim PP(\lambda(t))$, where $\lambda(t)$ is the rate function. Then the following properties hold:

1. $N(0) = 0$.
2. For any $s, t \in \mathbb{R}^+$, $(N(s+t) - N(s)) \perp (N(s))$. In other words, memorylessness still applies.
3. For any $s \in \mathbb{R}^+$, $N(s+t) - N(s) \sim \text{Poisson}\left(\int_s^{s+t} \lambda(x)\, dx\right)$.

Notice that the homogeneous PP is just a special case of the inhomogeneous PP; if the rate function is constant, then we get the three properties of the homogeneous PP.

4.3 Poisson Point Process

How can we generalize the Poisson process? The PP is a model of how many times something occurs within a specific time frame. To generalize, we ask how we can model the number of occurrences within a specific frame? Think: can we model how many stars there are within a specific window?

Definition 4.7. Let $N(T) \sim PPP(\lambda(T))$, where $\lambda(\cdot)$ is a measure. Then the following properties hold:

1. $N(\emptyset) = 0$.
2. If $A \cap B = \emptyset$, then $N(A) \perp N(B)$.
3. $N(A) \sim \text{Poisson}(\lambda(A))$.

5 Multivariate Gaussian / Normal

Definition 5.1 (Multi-dimensional expectation and variance)**.** If $X = (X_1, ..., X_p) \in \mathbb{R}^p$ is a p-dimensional random vector, then define

$$\mathbb{E}(X) = (\mathbb{E}(X_1), ..., \mathbb{E}(X_p)) \in \mathbb{R}^p, \text{ and } \text{Cov}(X) = \mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T] \in \mathbb{R}^{p \times p}.$$

The diagonal of the covariance matrix gives the variances.

Theorem 5.2. *Suppose X is a p-dimensional random vector. Given fixed $A \in \mathbb{R}^{q \times p}$, and $b \in \mathbb{R}^q$,*

$$\mathbb{E}(AX + b) = A\mathbb{E}(X) + b, \text{ and } \text{Cov}(AX + b) = A\text{Cov}(X)A^T.$$

Proof. These are both easily shown by writing out all the components. The matrix proof of the covariance bit is

$$\begin{aligned} \text{Cov}(AX + b) &= \mathbb{E}[[AX + b - \mathbb{E}(AX + b)][AX + b - \mathbb{E}(AX + b)]^T] \\ &= \mathbb{E}[(AX - A\mathbb{E}(X))(AX - A\mathbb{E}(X))^T] \\ &= A\mathbb{E}[(X - \mathbb{E}(X))(X - \mathbb{E}(X))^T]A^T = A\text{Cov}(X)A^T. \end{aligned}$$

□

Definition 5.3. Suppose $X \in \mathbb{R}^p \sim N(\mu, \Sigma)$, where $\mu \in \mathbb{R}^p$ and $\Sigma \in \mathbb{R}^{p \times p}$. The density function of the multivariate Gaussian is

$$p(X) = (2\pi)^{-\frac{p}{2}} \cdot [\det(\Sigma)]^{-\frac{1}{2}} \cdot e^{-\frac{1}{2} \cdot (X-\mu)^T \Sigma^{-1} (X-\mu)}.$$

As expected, $\mathbb{E}(X) = \mu$ and $\text{Cov}(X) = \Sigma$.

Proposition 5.4. *Suppose $X \in \mathbb{R}^p \sim N(\mu, \Sigma)$. Then linear combinations of X are multivariate Gaussian also. In particular, if $A \in \mathcal{L}(\mathbb{R}^p, \mathbb{R}^q) = \mathbb{R}^{q \times p}$, then*

$$AX \sim N(A\mu, A\Sigma A^T).$$

Proposition 5.5. *If $X \in \mathbb{R}^q, Y \in \mathbb{R}^p$ are jointly Gaussian (i.e. multivariate Gaussian), and $\text{Cov}(X, Y) = 0$, then $X \perp Y$. In other words, given two random vectors that are jointly distributed to the multivariate Gaussian, independence is equivalent to covariance 0.*

Proof. We want to show that $p(x,y) = p(x)p(y)$. Since X, Y are jointly Gaussian, we have

$$\begin{bmatrix} X \\ Y \end{bmatrix} = N\left(\begin{bmatrix} \mu_X \\ \mu_Y \end{bmatrix}, \begin{bmatrix} \Sigma_{XX} & \Sigma_{XY} \\ \Sigma_{YX} & \Sigma_{YY} \end{bmatrix} \right).$$

Since $\text{Cov}(X,Y) = 0$, Σ_{XY} and Σ_{YX} are both zero. By writing out the density and factoring, we get our result. □

6 Chi-Square, T

Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, \sigma^2)$. Recall that $\hat{\mu} = \bar{X} \sim N(\mu, \frac{\sigma^2}{n})$ exactly (since all the iid variables are normal also). We want to construct a confidence interval for μ.

1. **If σ^2 is known,** then since we know the exactly distribution of $\hat{\mu}$, we can construct the exact $(1-\alpha)$ confidence interval

$$\mu \in \left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right].$$

2. **If σ^2 is unknown,** we can use Wald's method. Setting $\hat{\sigma} = \frac{1}{n}\sum(X_i - \bar{X})^2$, since $\hat{\sigma} \to \sigma$, we have from Slutsky's Theorem that

$$\frac{\sqrt{n}(\hat{\mu} - \mu)}{\hat{\sigma}} \to N(0,1).$$

   Hence the approximate $(1-\alpha)$ confidence interval is

$$\mu \in \left[\bar{X} - z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + z_{1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right].$$

   But if we try to use Wilson's or VST, we run into a problem. Since the normal distribution is made up of two parameters, we cannot solve the equations required by Wilson's or VST. It turns out that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2}} \sim_{\text{exact}} t_{n-1}.$$

Definition 6.1. Let $Z_1, ..., Z_n \sim_{\text{i.i.d.}} N(0,1)$. Then

$$Y = Z_1^2 + ... + Z_n^2 \sim \chi_n^2.$$

Note that $\mathbb{E}(Y) = n$ and $\text{Var}(Y) = 2n$.

Remark 6.2. The χ_n^2 distribution is also a special case of the Gamma distribution. Specifically, $\chi_n^2 \sim \text{Gamma}(\frac{n}{2}, 2)$.

Definition 6.3. Let $X, Z_1, ..., Z_n \sim_{\text{i.i.d.}} N(0,1)$. Then $Y = \sum_{i=1}^n Z_i^2 \sim \chi_n^2$. Then

$$\frac{X}{\sqrt{\frac{Y}{n}}} \sim t_n.$$

In fact, as long as $X \perp Y$ and $Y \sim \chi_n^2$ and $X \sim N(0,1)$, the expression above follows exactly the t_n distribution.

Remarks 6.4. 1. Since $\mathbb{E}(Y) = n$, by the Law of Large Numbers $\frac{Y}{n} \to 1$. Hence $\frac{X}{\sqrt{Y/n}} \to N(0,1)$ as $n \to \infty$, i.e. the t-distribution converges in distribution to the standard normal.

2. The $t_1 = \frac{X}{|Z_1|}$ distribution (t distribution with 1 degree of freedom) is called the Cauchy distribution.

   - The Cauchy distribution has undefined expectation and variance. This is because the tails are so fat that the integrals for the expectation + variance do not exist.
   - The sample mean of iid Cauchy RVs is distributed according to the Cauchy distribution i.e. $X_1, ..., X_n \sim_{\text{i.i.d.}}$ Cauchy $\implies \bar{X} \sim$ Cauchy. Contrast this with $X_1, ..., X_n \sim_{\text{i.i.d.}} N(0,1) \implies \bar{X} \sim N(0, \frac{1}{n})$.

Theorem 6.5. *Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, \sigma^2)$. Then the following hold:*

1. ***(CLT)*** $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$.
2. ***(Chi-square)*** $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$.
3. ***(Independence)*** $\bar{X} \perp \sum_{i=1}^n (X_i - \bar{X})^2$.
4. ***(T-distribution)*** $\frac{\sqrt{n}(\bar{X}-\mu)}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2}} \sim t_{n-1}$.

Proof of (1). This follows immediately from the CLT. □

Proof of (2). First, notice that for any $X \sim N(\mu, \sigma^2)$, $X = \mu + \sigma Z$, where $Z \sim N(0,1)$. This also implies that $\bar{X} = \mu + \sigma \bar{Z}$, where $\bar{Z} = \frac{Z_1 + ... + Z_n}{n}$ and $Z_1, ..., Z_n \sim_{\text{i.i.d.}} N(0,1)$. Hence,

$$\begin{aligned} \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} &= \frac{\sum_{i=1}^n ((\mu + \sigma Z_i) - (\mu + \sigma \bar{Z}))^2}{\sigma^2} \\ &= \sum_{i=1}^n (Z_i - \bar{Z})^2. \end{aligned}$$

Hence it suffices to just use standard normals. We now prove by induction. If $n = 2$, we have

$$\sum_{i=1}^2 (Z_i - \bar{Z})^2 = \left(\frac{Z_1 - Z_2}{\sqrt{2}} \right)^2.$$

Since $\frac{Z_1 - Z_2}{\sqrt{2}} \sim N(0,1)$, we get our result for $n = 2$. Next, assume result (2) holds for m. Denote $\bar{Z}_m = \frac{Z_1 + ... + Z_m}{m}$ and $\bar{Z}_{m+1} = \frac{Z_1 + ... + Z_{m+1}}{m+1}$. Then

$$\bar{Z}_{m+1} = \frac{m\bar{Z}_m + Z_{m+1}}{m} = \frac{m}{m+1} \bar{Z}_m + \frac{1}{m+1} Z_{m+1}.$$

Now, we have

$$\begin{aligned}
&\sum_{i=1}^{m+1}(Z_i - \bar{Z}_{m+1})^2 \\
&= \sum_{i=1}^{m}(Z_i - \bar{Z}_{m+1})^2 + (Z_{m+1} - \bar{Z}_{m+1})^2 \\
&= \sum_{i=1}^{m}(Z_i - \bar{Z}_m + \bar{Z}_m - \bar{Z}_{m+1})^2 + (Z_{m+1} - \bar{Z}_{m+1})^2 \\
&= \sum_{i=1}^{m}(Z_i - \bar{Z}_m)^2 + m(\bar{Z}_m - \bar{Z}_{m+1})^2 + 2\sum_{i=1}^{m}(Z_i - \bar{Z}_m)(\bar{Z}_m - \bar{Z}_{m+1}) + (Z_{m+1} - \bar{Z}_{m+1})^2.
\end{aligned}$$

We see that

$$\begin{aligned}
\sum_{i=1}^{m}(Z_i - \bar{Z}_m)(\bar{Z}_m - \bar{Z}_{m+1}) &= (\bar{Z}_m - \bar{Z}_{m+1})\sum_{i=1}^{m}(Z_i - \bar{Z}_m) \\
&= (\bar{Z}_m - \bar{Z}_{m+1})(Z_1 + ... + Z_m - m\bar{Z}_m) = 0.
\end{aligned}$$

Now, notice that $\bar{Z}_{m+1} = \frac{m\bar{Z}_m + Z_{m+1}}{m+1}$. Therefore,

$$\begin{aligned}
\bar{Z}_m - \bar{Z}_{m+1} = \bar{Z}_m - \frac{m\bar{Z}_m + Z_{m+1}}{m+1} = \frac{1}{m+1}\cdot(\bar{Z}_m - Z_{m+1}), \text{ and} \\
Z_{m+1} - \bar{Z}_{m+1} = Z_{m+1} - \frac{m\bar{Z}_m + Z_{m+1}}{m+1} = \frac{m}{m+1}\cdot(Z_{m+1} - \bar{Z}_m).
\end{aligned}$$

Hence, it follows that

$$m(\bar{Z}_m - \bar{Z}_{m+1})^2 + (Z_{m+1} - \bar{Z}_{m+1})^2 = \frac{m}{m+1}(\bar{Z}_m - Z_{m+1})^2 = \left(\sqrt{\frac{m}{m+1}}\bar{Z}_m - Z_{m+1}\right)^2.$$

Because $\bar{Z}_m \sim N(\mu, \frac{\sigma^2}{m})$ and $Z_{m+1} \sim N(\mu, \sigma^2)$, $\sqrt{\frac{m}{m+1}}\bar{Z}_m - Z_{m+1}$ is also normally distributed with mean 0 and variance 1. Furthermore, both $\bar{Z}_m$ and Z_{m+1} are independent from $\sum_{i=1}^{m}(Z_i - \bar{Z}_m)^2$: the former by part 3 of the theorem, and the latter by not appearing in the expression at all. Therefore, by the induction hypothesis, we have

$$\sum_{i=1}^{m+1}(Z_i - \bar{Z}_{m+1})^2 = \underbrace{\sum_{i=1}^{m}(Z_i - \bar{Z}_m)^2}_{\sim\chi^2_m} + \underbrace{\left(\sqrt{\frac{m}{m+1}}\bar{Z}_m - Z_{m+1}\right)^2}_{\sim\chi^2_1} \sim \chi^2_{m+1}, \text{ as desired.}$$

□

Proof of (3). We aim to prove the stronger claim that

$$\bar{X} \perp \begin{bmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}$$

.

First, because it is the linear combination of $(X_1, ..., X_n)^T$, notice that

$$\begin{bmatrix} \bar{X} \\ X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix} \text{ is a } (n+1)\text{-dimensional Gaussian.}$$

This is because all the entries in the random vector are linear combinations of $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, \sigma^2)$. (see Wikipedia). Next, we have

$$\text{Cov}\left(\bar{X}, \begin{bmatrix} \bar{X} \\ X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}\right) = \begin{bmatrix} \text{Cov}(\bar{X}, X_1 - \bar{X}) \\ \vdots \\ \text{Cov}(\bar{X}, X_n - \bar{X}). \end{bmatrix}$$

See that $\text{Cov}(\bar{X}, X_1 - \bar{X}) = \text{Cov}(\bar{X}, \bar{X}) - \text{Cov}(\bar{X}, X_1)$. First, $\text{Cov}(\bar{X}, \bar{X}) = \text{Var}(\bar{X}) = \frac{\sigma^2}{n}$. Second, we have

$$\begin{aligned} \text{Cov}(\bar{X}, X_i) &= \text{Cov}\left(\frac{X_1 + ... + X_n}{n}, X_i\right) \\ &= \frac{\text{Cov}(X_1, X_i) + ... + \text{Cov}(X_n, X_i)}{n} = \frac{\sigma^2}{n}. \end{aligned}$$

Therefore,

$$\text{Cov}\left(\bar{X}, \begin{bmatrix} \bar{X} \\ X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}\right) = 0, \text{ so } \bar{X} \perp \begin{bmatrix} X_1 - \bar{X} \\ \vdots \\ X_n - \bar{X} \end{bmatrix}.$$

□

Proof of (4). By (1) and (2), we have that $\frac{\sqrt{n}(\bar{X}-\mu)}{\sigma} \sim N(0,1)$ and $\frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2} \sim \chi^2_{n-1}$. Because (1)'s expression is a function of $\bar{X}$, and (2)'s expression is a function of $\sum_{i=1}^n (X_i - \bar{X})^2$, it follows that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sigma} \perp \frac{\sum_{i=1}^n (X_i - \bar{X})^2}{\sigma^2}.$$

By definition of t distribution, it follows after cancelling the σs that

$$\frac{\sqrt{n}(\bar{X} - \mu)}{\sqrt{\frac{1}{n-1} \cdot \sum_{i=1}^n (X_i - \bar{X})^2}} \sim t_{n-1}.$$

This allows us to come up with an exact confidence interval for μ when σ^2 is unknown. □

7 Hypothesis Testing

Definition 7.1. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} p(\theta)$, where $p(\theta)$ is some distribution with $\theta \in \mathbb{R}^k$. Then a hypothesis test consists of

1. a null and alternative hypothesis, H_0 and H_1 respectively,

2. a test statistic $T = T(X_1, ..., X_n)$ that is a function of the data,

3. a rejection region R, and

4. a testing procedure that rejects H_0 if $T \in R$.

Definition 7.2. The Type I error is $\mathbb{P}(T \in R \mid H_0 \text{ is true})$, and the Type II error is $\mathbb{P}(T \notin R \mid H_1 \text{ is true})$. The power is $\mathbb{P}(T \in R \mid H_1 \text{ is true})$.

The goal is to find a test such that the Type I error $\leq \alpha$, and the Type II error is small as possible.

We begin with two examples.

Example 7.3. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, 1)$, where μ is unknown. We set

$$H_0 : \mu = 0 \text{ and } H_1 : \mu > 0.$$

Use $\bar{X}$ as the test statistic. We want to reject the null if $\bar{X}$ is too large i.e. $\bar{X} > c$ for some threshold c. The goal is to set this threshold such that $\mathbb{P}(\bar{X} > c \mid H_0) = \alpha$. Under H_0, we have that $\sqrt{n}\bar{X} \sim N(0, 1)$, so

$$\mathbb{P}(\bar{X} > c \mid H_0) = \alpha \iff c = \frac{z_{1-\alpha}}{\sqrt{n}}.$$

Example 7.4. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, 1)$, where μ is unknown. We set

$$H_0 : \mu = 0 \text{ and } H_1 : \mu \neq 0.$$

Use $\bar{X}$ as the test statistic. We want to reject the null if $\bar{X}$ is too large or too small i.e. $|\bar{X}| > c$ for some threshold c. The goal is to set this threshold such that $\mathbb{P}(|\bar{X}| > c \mid H_0) = \alpha$. Under H_0, we have that $\sqrt{n}\bar{X} \sim N(0, 1)$, so

$$\mathbb{P}(|\bar{X}| > c \mid H_0) = \alpha \iff c = \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}.$$

Example 7.5. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, \sigma^2)$, where μ and σ^2 are unknown. We set

$$H_0 : \mu = 0 \text{ and } H_1 : \mu > 0.$$

Since σ^2 is unknown, use the t-statistic instead for the test statistic, since we know that

$$T = \frac{\sqrt{n}\bar{X}}{\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2}} \sim t_{n-1}.$$

We arrive at the conclusion by similar logic:

$$\mathbb{P}(T > c \mid H_0) = \alpha \iff c = t_{n-1,1-\alpha}.$$

If instead $H_1 : \mu \neq 0$, then using $c = t_{n-1,1-\frac{\alpha}{2}}$ will ensure that the Type I error is α.

If the null is composite, then the idea is to find the rejection region by controlling for the worst possible case under the null. In other words, if the set of parameters given by the null is N, then we want to find a rejection region R such that

$$\sup_{\theta \in N} \mathbb{P}(T_\theta \in R) \leq \alpha, \text{ where } T_\theta \text{ is a test statistic dependent on } \theta.$$

By controlling for the worst possible case, we effectively turn the composite case into a simple one.

Example 7.6 (Composite null)**.** Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, 1)$, and $H_0 : \mu \leq 0 \mid H_1 : \mu > 0$. Then we want to find a threshold c such that

$$\sup_{\mu \leq 0} \mathbb{P}(\bar{X}_\mu > c) = \alpha.$$

Given $\mu \leq 0$, $\sqrt{n}(\bar{X}_\mu - \mu) \sim N(0, 1)$. Hence we have

$$\begin{aligned}\sup_{\mu \leq 0} \mathbb{P}(\bar{X}_\mu > c) &= \sup_{\mu \leq 0} \mathbb{P}(\sqrt{n}(\bar{X} - \mu) > \sqrt{n}c - \sqrt{n}\mu) \\ &= \sup_{\mu \leq 0} \mathbb{P}(N(0, 1) > \sqrt{n}c - \sqrt{n}\mu).\end{aligned}$$

Because we are only consider nonpositive μ, the probability is greatest when $\mu = 0$. Therefore, we want c such that

$$\sup_{\mu \leq 0} \mathbb{P}(\bar{X}_\mu > c) = \mathbb{P}(N(0, 1) > \sqrt{n}c) = \alpha.$$

Hence, reject the null if $\bar{X} > \frac{z_{1-\alpha}}{\sqrt{n}}$.

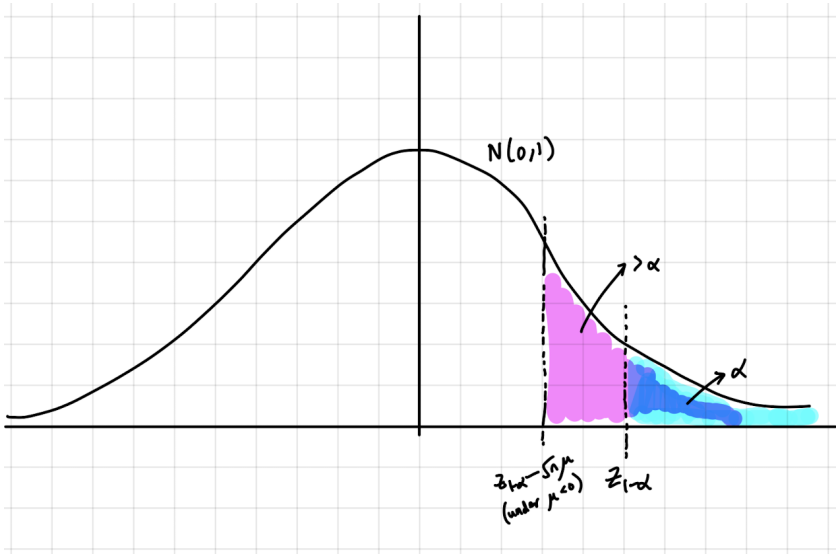
7.1 Power

The power of a hypothesis test is defined as $\mathbb{P}(T \in R \mid H_1)$ i.e. the probability of T falling in the rejection region given that the alternative is true. Because T is dependent on the hypothesis, we set $T_\theta = T$. Hence, it's possible to define $\mathbb{P}(T_\theta \in R)$ as a function of θ, where θ is either in the null or the alternative. This means that the Type I error can be interpreted as a special case of power.

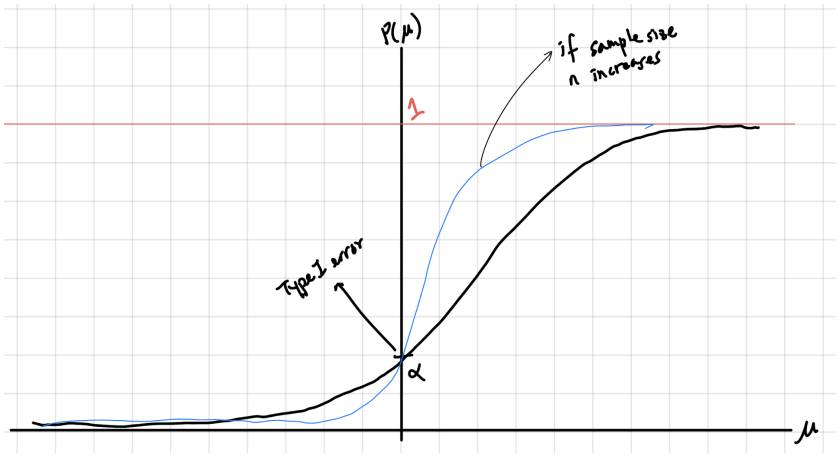
Example 7.7. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, 1)$, and $H_0 : \mu = 0 \mid H_1 : \mu > 0$. Reject the null if $\sqrt{n}\bar{X} > z_{1-\alpha}$. Since $\sqrt{n}\bar{X} \sim N(\sqrt{n}\mu, 1)$ where $\mu > 0$ under the alternative, the power is

$$\begin{aligned}
\mathbb{P}(\sqrt{n}\bar{X} > z_{1-\alpha}) &= \mathbb{P}(N(\sqrt{n}\mu, 1) > z_{1-\alpha}) \\
&= \mathbb{P}(N(0,1) + \sqrt{n}\mu > z_{1-\alpha}) \\
&= \mathbb{P}(N(0,1) > z_{1-\alpha} - \sqrt{n}\mu).
\end{aligned}$$

From the figure below, it is clear that the probability changes as a function of μ.



Hence, we can express the power as a function of μ i.e. $P(\mu) = \mathbb{P}(N(0,1) > z_{1-\alpha} - \sqrt{n}\mu)$. Notice that as we increase the sample size n, the movement of $z_{1-\alpha} - \sqrt{n}\mu$ becomes more sensitive for every movement of μ. Therefore, as $n \to \infty$, the power function becomes a step function, which is intuitive because we expect the hypothesis test to work 100% of the time with an infinite amount of data points.



7.2 Duality of Confidence Intervals and Hypothesis Tests

Confidence intervals are random intervals that have a $1-\alpha$ probability of containing the true parameter θ. In hypothesis testing, we define the rejection region such that the probability of falling inside the rejection region is less than α, given the null hypothesis. This means that under the null, we are effectively given a "true" parameter, and we design a rejection region that is **based on this "true" parameter**.

Furthermore, confidence intervals should have $1-\alpha$ coverage, but we aim to make them as small as possible. Similarly, hypothesis tests should have less than α Type I error, but we aim to make the Type II error as small as possible also.

Example 7.8. Suppose $X_1, ..., X_n \sim_{\text{i.i.d.}} N(\mu, 1)$, and $H_0 : \mu = \mu^* \mid H_1 : \mu \neq \mu^*$. Then under the null, $\sqrt{n}(\bar{X} - \mu^*) \to N(0,1)$. Hence, reject the null if $|\sqrt{n}(\bar{X} - \mu^*)| > z_{1-\frac{\alpha}{2}}$. The Type I error is

$$\mathbb{P}(|\sqrt{n}(\bar{X} - \mu^*)| > z_{1-\frac{\alpha}{2}}) = \alpha \iff \mathbb{P}(|\sqrt{n}(\bar{X} - \mu^*)| \leq z_{1-\frac{\alpha}{2}}) = 1 - \alpha.$$

This means that under the null, the interval $\left[\bar{X} - \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}, \bar{X} + \frac{z_{1-\frac{\alpha}{2}}}{\sqrt{n}}\right]$ has $1-\alpha$ coverage of μ^*. This resembles the confidence interval of μ^*, because through the null, we've effectively been given a "true" parameter.

Example 7.9. Suppose $X_1, .., X_n \sim_{\text{i.i.d.}} \text{Poisson}(\lambda)$, and $H_0 : \lambda = \lambda^* \mid H_1 : \lambda \neq \lambda^*$. Under the null, $\frac{\sqrt{n}(\bar{X} - \lambda^*)}{\sqrt{\lambda^*}} \to N(0,1)$. Hence reject the null if $\left|\frac{\sqrt{n}(\bar{X} - \lambda^*)}{\sqrt{\lambda^*}}\right| > z_{1-\frac{\alpha}{2}}$. The Type I error is

$$\mathbb{P}\left(\left|\frac{\sqrt{n}(\bar{X} - \lambda^*)}{\sqrt{\lambda^*}}\right| > z_{1-\frac{\alpha}{2}}\right) \approx \alpha \iff \mathbb{P}\left(\left|\frac{\sqrt{n}(\bar{X} - \lambda^*)}{\sqrt{\lambda^*}}\right| \leq z_{1-\frac{\alpha}{2}}\right) \approx 1 - \alpha.$$

This resembles the CI obtained from Wilson's method, since the denominator contains the "true" parameter λ^* under the null.

7.3 p-value

As far as convenience is concerned, you have to keep track of several numbers when doing a hypothesis test: the α level and the rejection region threshold(s). Wouldn't it just be easier to combine the two i.e. reject the null when the test statistic is less than α?

Theorem 7.10. *Suppose X is a random variable with CDF $F(t) = \mathbb{P}(X \leq t)$. Let $Y = F(X)$. Then $Y \sim \text{Uniform}[0,1]$.*

Proof. $\mathbb{P}(Y \le t) = \mathbb{P}(F(X) \le t) = \mathbb{P}(X \le F^{-1}(t)) = F(F^{-1}(t)) = t$, so $Y \sim \text{Uniform}[0,1]$. □

Example 7.11. Suppose $X_1, ..., X_n \sim N(\mu, 1)$ with $H_0 : \mu = 0 \mid H_1 : \mu > 0$. We reject the null if $\sqrt{n}\bar{X} > z_{1-\alpha}$. Note that $\sqrt{n}\bar{X} \sim N(0,1)$. Therefore, if F is the CDF of the standard normal, then we reject the null if $F(-\sqrt{n}\bar{X}) < \alpha$. In other words,

$$\begin{aligned}\sqrt{n}\bar{X} > z_{1-\alpha} &\iff -\sqrt{n}\bar{X} < z_\alpha = F^{-1}(\alpha) \\ &\iff F(-\sqrt{n}\bar{X}) < \alpha.\end{aligned}$$

Example 7.12. Suppose $X_1, ..., X_n \sim N(\mu, 1)$ with $H_0 : \mu = 0 \mid H_1 : \mu \neq 0$. We reject the null if $|\sqrt{n}\bar{X}| > z_{1-\frac{\alpha}{2}}$. Note that $\sqrt{n}\bar{X} \sim N(0,1)$. Define $F(t) = \mathbb{P}(-|\sqrt{n}\bar{X}| \le t)$. Then we reject the null if $F(-|\sqrt{n}\bar{X}|) < \alpha$. In other words,

$$-|\sqrt{n}\bar{X}| > z_{1-\frac{\alpha}{2}} = F^{-1}(\alpha) \iff F(-|\sqrt{n}\bar{X}|) < \alpha.$$

*Multiple Testing

Suppose we have the following scenario: you have n drugs, and you want to test the significance of each drug. For the ith drug, you construct a p-value p_i. Then, since p-values are defined to be uniform only under the null, then the ith null hypothesis is $H_{0i} : p_i \sim \text{Uniform}[0,1]$.

We have two questions:

1. **Is there a significant drug?** This is a binary question, and we can approach this question with the traditional hypothesis test.
2. **Which ones are significant?** This is non-binary, so we'll have to come up with another formulation of hypothesis testing.

Is there a significant drug?

We can construct a global null:

$$H_0 : \bigcap_{i=1}^{n} H_{0i} = \text{ every } p_i \sim \text{Uniform}[0,1].$$

This way, a rejection of the null means that there exists a significant drug. There are two possible tests:

- **Bonferroni test (no conditions).** Reject the null if there exists i such that $p_i < \frac{\alpha}{n}$.

  We show that this test does control the Type I error. Recall that $\mathbb{P}\left(\bigcup_{i=1}^{n} A_i\right) \le \sum_{i=1}^{n} \mathbb{P}(A_i)$. We have

$$\begin{aligned}\mathbb{P}(\text{Reject the null} \mid H_0) &= \mathbb{P}(\exists i \text{ s.t. } p_i < \frac{\alpha}{n} \mid H_0) \\ &= \mathbb{P}(p_0 < \frac{\alpha}{n} \text{ or ... or } p_n < \frac{\alpha}{n} \mid H_0) \\ &\le \sum_{i=1}^{n} \mathbb{P}(p_i < \frac{\alpha}{n} \mid H_0) \\ &\le \sum_{i=1}^{n} \frac{\alpha}{n} = \alpha \qquad \text{(because } p_i \sim \text{Uniform[0,1] under null)}\end{aligned}$$

- **Simes test (independence assumed).** If $p_1, ..., p_n$ are independent, then reject the null if

$$\min_{1 \le i \le n} \frac{np_{(i)}}{i} \le \alpha, \text{ where } p_{(i)} \text{ is the } i\text{th order statistic.}$$

The Simes test is based on the Simes theorem: If $U_1, ..., U_n \sim_{\text{i.i.d.}} \text{Uniform}[0,1]$, then

$$\min_{1 \le i \le n} \frac{nU_{(i)}}{i} \sim \text{Uniform}[0,1].$$

To show that this test controls the Type I error, we have from the Simes theorem that

$$\mathbb{P}(\text{Reject the null} \mid H_0) = \mathbb{P}\left(\min_{1 \le i \le n} \frac{np_{(i)}}{i} \le \alpha \mid H_0 \right) \le \alpha.$$

With the extra condition of independence, the Simes test is actually more powerful than the Bonferroni test. Why? The rejection condition gives

$$\begin{aligned} \min_{1 \le i \le n} \frac{np_{(i)}}{i} \le \alpha &\iff \exists i \text{ s.t. } \frac{np_{(i)}}{i} \\ &\iff \exists i \text{ s.t. } p_{(i)} \le \frac{\alpha \cdot i}{n}. \end{aligned}$$

The "rejection region" isn't as restricted as Bonferroni's, so Simes does have greater power.

Which drugs are significant?

We need a way to re-construct the structure of the hypothesis test from variability, since there are literally 2^n possible binary hypotheses we could make. Construct two sets $I_0, I_1 \subset \{1, ..., n\}$

$$\begin{aligned} I_0 &= \{i \mid H_{0i} \text{ is true}\} \\ I_1 &= \{i \mid H_{0i} \text{ is not true}\}. \end{aligned}$$

In particular, I_0, I_1 reflect which drugs are actually significant, whereas our rejection (or non-rejection) aims to find I_0, I_1. We summarize this in our table:

	reject	no reject
I_0	false rejection	
I_1		false acceptance

In the traditional hypothesis test, we aim to control the Type I error, and try to keep the Type II error as low as possible. Analogously, we need to aim to control the "false rejection" numbers, and keep the "false acceptance" numbers as low as possible. This leads to an analogous notion of the Type I error: the family-wise error rate (FWER).

Definition 7.13. The family-wise error rate is an analog of Type I error for multiple testing. It is

$$\begin{aligned} \text{FWER} &= \mathbb{P}(\exists \text{false discovery}) \\ &= \mathbb{P}(\exists i \in I_0 \text{ s.t. } H_{0i} \text{ is rejected}). \end{aligned}$$

The goal is then to control the FWER.

- **Bonferroni test (no conditions)** Reject H_{0i} if $p_i < \frac{\alpha}{n}$.

This test controls the FWER:

$$\begin{aligned}
\text{FWER} &= \mathbb{P}(\exists i \in I_0 \text{ s.t. } H_{0i} \text{ is rejected}) \\
&= \mathbb{P}(\exists i \in I_0 \text{ s.t. } p_i < \frac{\alpha}{n}) \\
&\leq \sum_{i \in I_0} \mathbb{P}(p_i < \frac{\alpha}{n}) \\
&= \frac{\#(I_0) \cdot \alpha}{n} \leq \alpha. \qquad (I_0 \text{ is exactly the set where } p_i \sim \text{Uniform}[0,1])
\end{aligned}$$

- **Holm test (no conditions).**
- **Hochberg test (independence assumed).**

The FWER is pretty restrictive, however. Controlling for it means controlling the probability of any *false discovery at all*. From the table, define V, V_1, R

	reject	no reject
I_0	V	
I_1	V_1	

and $R = V + V_1$.

Definition 7.14 (False discovery rate (FDR))**.** A less stringent analog to Type I error than FWER, the FDR is defined

$$\text{FDR} = \mathbb{E}\left(\frac{V}{\max(R,1)}\right).$$

Proposition 7.15. FDR $\leq$ FWER*. In other words, FWER is more stringent than FDR.*

Proof.

$$\begin{aligned}
\text{FDR} &= \mathbb{E}\left(\frac{V}{\max(R,1)}\right) \\
&= \mathbb{E}\left(\frac{V}{\max(R,1)}(\mathbb{1}_{V>0} + \mathbb{1}_{V=0})\right) \\
&= \mathbb{E}\left(\frac{V}{\max(R,1)} \cdot \mathbb{1}_{V>0}\right) + \mathbb{E}\left(\underbrace{\frac{V}{\max(R,1)} \cdot \mathbb{1}_{V=0}}_{=0}\right) \\
&= \mathbb{E}\left(\underbrace{\frac{V}{R}}_{\leq 1} \cdot \mathbb{1}_{V>0}\right) \leq \mathbb{E}(\mathbb{1}_{V>0}) = \mathbb{P}(V>0) = \text{FWER}.
\end{aligned}$$

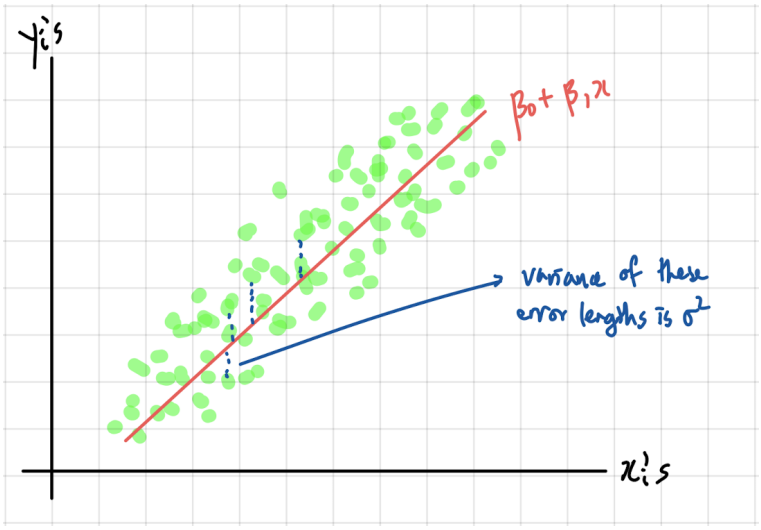
□

One test for multiple testing using FDR is the Benjamin-Hochberg procedure: order independent $p_1, ..., p_n$ into $p_{(1)} \leq ... \leq p_{(n)}$. The first time $p_{(j)} \leq \frac{j \cdot \alpha}{n}$, reject $H_{0(1)}, ..., H_{0(j)}$. It can be shown that under independence, we can control the FDR i.e. FDR $\leq \alpha$.

8 Linear Regression

8.1 Univariate Linear Regression

Suppose we have independent $y_i \sim N(\beta_0 + \beta_1 x_i, \sigma^2)$ for $i = 1, ..., n$. Note that this is equivalent to saying that for $y_i = \beta_0 + \beta_1 x_i + \sigma z_i$, where $z_i \sim_{\text{i.i.d.}} N(0, 1)$. Keep in mind that here, the x_i's are not random.



Proposition 8.1. *The* MLE*s for* (β_0, β_1) *are*

$$\hat{\beta}_0 = \bar{y} - \hat{\beta}_1 \bar{X}, \text{ and}$$
$$\hat{\beta}_1 = \frac{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})(y_i - \bar{y})}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Proof. Follows from painful algebra-crunching. Something to note is that maximizing the likelihood in this case is equivalent to minimizing $\sum_{i=1}^n (y_i - \beta_0 - \beta_1 x_i)^2$ i.e. minimizing the least squared error. □

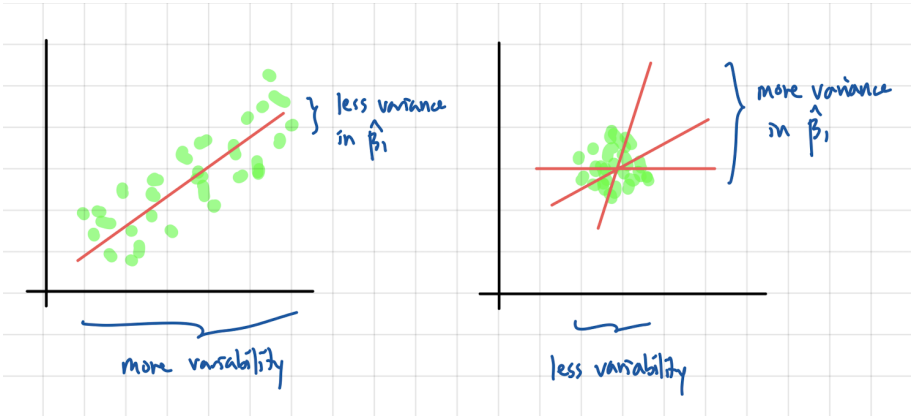
Proposition 8.2. *The* MLE*s* $(\hat{\beta}_0, \hat{\beta}_1)$ *are unbiased estimators.*

Proposition 8.3. *The variances of* $(\hat{\beta}_0, \hat{\beta}_1)$ *are*

$$\text{Var}(\hat{\beta}_1) = \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sigma^2}{n} \cdot \frac{1}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}$$
$$\text{Var}(\hat{\beta}_0) = \frac{\sigma^2}{n} + \frac{\sigma^2 \cdot \bar{X}^2}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sigma^2}{n}\left(1 + \frac{\bar{X}^2}{\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2}\right).$$

Proof. Pure calculation. For $\text{Var}(\hat{\beta}_0)$, realize that $\text{Cov}(\bar{y}, \hat{\beta}_1) = 0$. □

Remark 8.4. Note that $\hat{\beta}_1$ gives the slope of the estimator line. We can an observation from the expression of $\text{Var}(\hat{\beta}_1)$. Looking at the expression, σ^2 is the noise level of the error terms, n is the sample size, and $\frac{1}{n}\sum_{i=1}^n (X_i - \bar{X})^2$ is the variability of the x's. This means that the more "spread out" the x's are, the smaller the variance of $\hat{\beta}_1$ will be. In other words, wider and more diverse data makes the slope easier to discern.



Proposition 8.5. *The covariance of* $\hat{\beta}_0, \hat{\beta}_1$ *is*

$$\text{Cov}(\hat{\beta}_0, \hat{\beta}_1) = -\frac{\sigma^2 \cdot \bar{X}}{\sum_{i=1}^n (X_i - \bar{X})^2}.$$

Theorem 8.6. *Suppose* $y_i \sim_{ind.} N(\beta_0 + \beta_1 x_i, \sigma^2)$ *for* $i = 1, ..., n$. *Then*

1. $\begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{bmatrix} \sim N\left(\begin{bmatrix} \beta_0 \\ \beta_1 \end{bmatrix}, \begin{bmatrix} \text{Var}(\hat{\beta}_1) & \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) \\ \text{Cov}(\hat{\beta}_0, \hat{\beta}_1) & \text{Var}(\hat{\beta}_0) \end{bmatrix}\right).$
2. $\dfrac{\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}{\sigma^2} \sim \chi^2_{n-2}.$
3. $\displaystyle\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2 \perp \begin{bmatrix} \hat{\beta}_1 \\ \hat{\beta}_0 \end{bmatrix}.$
4. *All kinds of* t *statistics, e.g.* $\dfrac{\hat{\beta}_1 - \beta_1}{\sqrt{\frac{1}{\sum_{i=1}^n (X_i - \bar{X})^2} \cdot \frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{\beta}_0 - \hat{\beta}_1 x_i)^2}} \sim t_{n-2}.$

Proof. Notice that $Y = [y_1, ..., y_n]^T$ is Gaussian. Since $\hat{\beta}_0, \hat{\beta}_1$ are both linear combinations of $y_1, ..., y_n$, they are joint Gaussian. This gives (1). Even further, any linear combination of $\hat{\beta}_0, \hat{\beta}_1$ is also joint Gaussian. This means that $Y^* = [\hat{\beta}_0 + \hat{\beta}_1 x_1, ..., \hat{\beta}_0 + \hat{\beta}_1 x_n]^T$ is also Gaussian. This gives the rest of the results.

Essentially, the Gaussian-ness of Y enables us to show all the results. □

Inference on the prediction Suppose we want to predict the response for some new x^*. Our candidate is $y = \beta_0 + \beta_1 x^*$, while our prediction is $y^* = \hat{\beta_0} + \hat{\beta_1} x^*$. Let's try to construct a confidence interval for y.

First, notice that y^* is Gaussian distributed, since $[\hat{\beta_1}, \hat{\beta_0}]^T$ is Gaussian. To compute the expectation and variance of y^*, we have

$$\begin{aligned}\mathbb{E}(y^*) &= \beta_0 + \beta_1 x^* \\ \text{Var}(y^*) &= \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}(x^* - \bar{X})^2.\end{aligned}$$

Therefore, $y^* \sim N\left(\beta_0 + \beta_1 x^*, \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}(x^* - \bar{X})^2\right)$. To construct a confidence interval for y, we divide into the familiar two cases on the knowability of σ^2:

1. If σ^2 is known, then use

$$y^* \sim N\left(\beta_0 + \beta_1 x^*, \frac{\sigma^2}{n} + \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\boxed{(x^* - \bar{X})^2}\right)$$

.

2. If σ^2 is unknown, then use

$$\frac{y^* - y}{\sqrt{\left(\frac{1}{n} + \frac{\boxed{(x^* - \bar{X})^2}}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)\left(\frac{1}{n-2}\sum_{i=1}^n (y_i - \hat{\beta_0} - \hat{\beta_1} x_i)^2\right)}} \sim t_{n-1}.$$

The critical insight is the boxed term $(x^* - \bar{X})^2$. Notice that the further x^* is from the sample mean, the wider the confidence interval gets. This makes sense: the area closest to the sample mean will have the densest amount of information, and thus more information is available to make the confidence interval smaller.

8.2 Multivariate Linear Regression

Suppose $y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1} x_{i,p-1} + \sigma z_i$, where $z_i \sim_{\text{i.i.d.}} N(0,1)$ for $i = 1, ..., n$. We can express these y_i's succinctly as

$$\underbrace{\begin{bmatrix} y_1 \\ \vdots \\ y_n \end{bmatrix}}_{y} = \underbrace{\begin{bmatrix} 1 & x_{1,1} & x_{1,2} & \cdots\cdots & x_{1,p-1} \\ \vdots & \vdots & \vdots & \vdots & \vdots \\ 1 & x_{n,1} & x_{n,2} & \cdots\cdots & x_{n,p-1} \end{bmatrix}}_{X} \underbrace{\begin{bmatrix} \beta_0 \\ \vdots \\ \beta_{p-1} \end{bmatrix}}_{\beta} + \sigma \underbrace{\begin{bmatrix} z_1 \\ \vdots \\ z_n \end{bmatrix}}_{z}, \text{ i.e. } y = X\beta + \sigma z.$$

Note that $y \sim N(X\beta, \sigma^2 \text{Id}_n)$. We can try to find the MLE for β. In particular, we have

$$\text{MLE} = \text{argmax}_\beta \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n e^{-\frac{1}{2\sigma^2}||y - X\beta||^2} = \text{argmin}_\beta ||y - X\beta||^2.$$

Theorem 8.7. *The projection $\hat{\beta} = (X^T X)^{-1} X^T y$ is the MLE/LSE for β. In particular, $\hat{\beta}$ gives the least possible value for $||y - X\beta||^2$ i.e. for any $\beta \in \mathbb{R}^p$, $||y - X\hat{\beta}||^2 \leq ||y - X\beta||^2$.*

Proof. Let $\beta \in \mathbb{R}^p$. We have

$$\begin{aligned}||y - X\beta||^2 &= ||y - X\hat{\beta} + X\hat{\beta} - X\beta||^2 \\ &= ||y - X\hat{\beta}||^2 + ||X\hat{\beta} - X\beta||^2 + 2\langle y - X\hat{\beta}, X\hat{\beta} - X\beta\rangle.\end{aligned}$$

It can be shown that $\langle y - X\hat{\beta}, X\hat{\beta} - X\beta\rangle = 0$, which makes sense since $y - X\hat{\beta}$ is orthogonal to $X(\hat{\beta} - \beta)$. Therefore,

$$||y - X\beta||^2 = ||y - X\hat{\beta}||^2 + ||X\hat{\beta} - X\beta||^2 \geq ||y - X\hat{\beta}||^2.$$

□

Theorem 8.8. $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$.

Proof. Since $y \sim N(X\beta, \sigma^2 I_n)$ is Gaussian, and $\hat{\beta}$ is just a linear transformation applied to y, it is also normally distributed. Now,

$$\begin{aligned}\mathbb{E}(\hat{\beta}) &= (X^TX)^{-1}X^T\mathbb{E}(y) = (X^TX)^{-1}X^T(X\beta) = \beta, \text{ and} \\ \text{Cov}(\hat{\beta}) &= (X^TX)^{-1}X^T\text{Cov}(y)(X^TX)^{-1}X^T \\ &= \sigma^2(X^TX)^{-1}X^TI_n(X^TX)^{-1}X^T \\ &= \sigma^2(X^TX)^{-1}.\end{aligned}$$

□

Lemma 8.9. *Suppose $Z \sim N(\boldsymbol{0}, I_n)$, and P is a projection matrix of rank r. Then $||PZ||^2 \sim \chi^2_r$.*

Proof. Considering the eigenvalue decomposition $P = U\Lambda U^T$ and setting $W = U^TZ$,

$$||PZ||^2 = (PZ)^T(PZ) = Z^TP^TPZ = Z^TPZ = (U^TZ)^T\Lambda(U^TZ) = W^T\Lambda W.$$

If $W^T = [w_1 \cdots w_n]$, then note that $W = U^TZ \sim N(\mathbf{0}, I_n)$, so

$$W^T\Lambda W = w_1^2 + ... + w_r^2, \text{ and } w_1, ..., w_r \sim_{\text{i.i.d.}} N(0, 1).$$

Therefore, $||PZ||^2 \sim \chi^2_r$. □

Theorem 8.10 (Multivariate Sampling Theorem)**.** *Suppose $y \sim N(X\beta, \sigma^2 I_n)$. Then the following hold:*

1. $\hat{\beta} \sim N(\beta, \sigma^2(X^TX)^{-1})$.
2. $\dfrac{||y - X\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-p}$.
3. *$\hat{\beta}$ and $y - X\hat{\beta}$ are independent.*
4. *All kinds of t-statistics.*

Proof of (2). Because $y \sim N(X\beta, \sigma^2 I_n)$, we can write $y = X\beta + \sigma Z$ where $Z \sim N(\mathbf{0}, I_n)$. Therefore

$$\begin{aligned}y - X\hat{\beta} &= y - X(X^TX)^{-1}X^Ty \\ &= (I_n - X(X^TX)^{-1}X^T)y \\ &= (I_n - X(X^TX)^{-1}X^T)(X\beta + \sigma Z) = (I_n - X(X^TX)^{-1}X^T)(\sigma Z).\end{aligned}$$

This means that

$$\frac{||y - X\hat{\beta}||^2}{\sigma^2} = ||(I_N - X(X^TX)^{-1}X^T)Z||^2.$$

We already know that $I_n - X(X^TX)^{-1}X^T$ is a projection matrix, so the lemma above gives that $||(I_n - X(X^TX)^{-1}X^T)Z||^2$ is chi-square distributed. It remains to find the degrees of freedom. We have

$$\begin{aligned}\mathrm{Tr}(I_n - X(X^TX)^{-1}X^T) &= \mathrm{Tr}(I_n) - \mathrm{Tr}(X(X^TX)^{-1}X^T) \\ &= n - \mathrm{Tr}((X^TX)^{-1}X^TX) = n - p.\end{aligned}$$

Hence $\frac{||y - X\hat{\beta}||^2}{\sigma^2} \sim \chi^2_{n-p}$.

□

Proof of (3). Since both $\hat{\beta}$ and $y - X\hat{\beta}$ are both linear combinations of y, $(\hat{\beta}, y - X\hat{\beta})$ is joint Gaussian. It then suffices to show that the covariance is zero. Therefore,

$$\begin{aligned}\mathrm{Cov}(\hat{\beta}, y - X\hat{\beta}) &= \mathrm{Cov}((X^TX)^{-1}X^Ty, y - X(X^TX)^{-1}X^Ty) \\ &= (X^TX)^{-1}X^T\mathrm{Cov}(y)(I_n - X(X^TX)^{-1}X^T)^T \\ &= \sigma^2(X^TX)^{-1}X^TI_n(I_n - X(X^TX)^{-1}X^T) \\ &= \sigma^2(X^TX)^{-1}(X^T - X^T) = 0.\end{aligned}$$

□

Proof of (4). Follows from (1), (2), and (3). □

8.3 Hypothesis Testing

Suppose our regular model $y \sim N(X\beta, \sigma^2 I_n)$. We want to test if any of the regression coefficients are significant. Fashion the hypotheses:

$$\begin{aligned}&H_0 : \beta_1 = \cdots = \beta_{p-1} = 0 \quad &&\iff y_i = \beta_0 + \sigma z_i, \text{ where } z_i \sim_{\text{i.i.d.}} N(0,1) \\ &H_1 : \text{otherwise} &&\iff y_i = \beta_0 + \beta_1 x_{i,1} + ... + \beta_{p-1}x_{i,p-1} + \sigma z_i, \text{ where } z_i \sim_{\text{i.i.d.}} N(0,1).\end{aligned}$$

To gain some intuition for these hypotheses, consider these exercises:

- Assume H_0 is true. Then $y_i \sim_{\text{i.i.d.}} N(\beta_0, \sigma^2)$, which means that the MLE/LSE of β_0 is $\bar{y}$. This is to say that $[\bar{y}, ..., \bar{y}]^T = \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^Ty$ is the best "fit" for $[y_1, ..., y_n]^T$ that the model under H_0 spits out.
- Assume H_1 is true. Then we calculated that

$$\hat{y} = X\hat{\beta} = \underbrace{X(X^TX)^{-1}X^T}_{=H}y \text{ is the best fit for } [y_1, ..., y_n]^T.$$

  Keep in mind that in any case, $\hat{\beta}$ gives the least distance from the column space of X and y.

The intuition is that if H_0 is true, then we expect $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^Ty$ and $\hat{y}$ to be "close" together. But if instead H_1 is true, then they should "far" apart.

Definition 8.11. Qualifications of variance depending on H_0 and H_1:

$$\text{TSS} = \sum_{i=1}^{n}(y_i - \bar{y})^2 = ||y - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y||^2,$$
$$\text{RSS} = \sum_{i=1}^{n}(y_i - \hat{y}_i)^2 = ||y - \hat{y}||^2,$$
$$\text{MSS} = \sum_{i=1}^{n}(\hat{y}_i - \bar{y})^2 = ||\hat{y} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y||^2.$$

In hypothesis testing, we start with assuming that the null is true, then see if the data gives us enough evidence to reject the null. In line with this reasoning and the intuition outlined above, take a look at the TSS. It tells how far apart y and $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y$ are. We can further break down the TSS with the following lemma.

Lemma 8.12. $\text{TSS} = \text{RSS} + \text{MSS}$.

Proof. Follows by recognizing that $HX = X(X^TX)^{-1}X^TX = X$. Can also be proved by noting that $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$ is in the column space of X, so we can just apply Theorem 8.7. □

Definition 8.13 (F-distribution)**.** Suppose $Y_1 \sim \chi^2_{d_1}$, $Y_2 \sim \chi^2_{d_2}$, and $Y_1 \perp Y_2$. Then

$$\frac{Y_1/d_1}{Y_2/d_2} \sim F_{d_1,d_2}.$$

Theorem 8.14. *Suppose $y \sim N(X\beta, \sigma^2 I_n)$. Then the following hold:*

1. $\text{TSS} = \text{RSS} + \text{MSS}$.
2. $\text{RSS} \perp \text{MSS}$
3. $\frac{\text{RSS}}{\sigma^2} \sim \chi^2_{n-p}$.
4. *Under H_0,* $\frac{\text{MSS}}{\sigma^2} \sim \chi^2_{p-1}$.
5. *Under H_0,* $\frac{\text{TSS}}{\sigma^2} \sim \chi^2_{n-1}$.
6. *Under H_0,* $\frac{\text{MSS}/(p-1)}{\text{RSS}/(n-p)} \sim F_{p-1,n-p}$.

Proof of (1). Follows from the lemma above. □

Proof of (2). Follows from showing that $y - \hat{y} \perp \hat{y} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y$ via showing that

$$\text{Cov}\left(y - \hat{y}, \hat{y} - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y\right) = 0.$$

□

Proof of (3). Follows from Theorem 8.10. □

Proof of (4). We have that

$$\begin{aligned}\frac{\text{MSS}}{\sigma^2} &= \frac{||(H - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)y||^2}{\sigma^2} \\ &= \frac{||(H - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y)(\beta_0\mathbb{1}_n + \sigma Z)||^2}{\sigma^2} && \text{(where } Z \sim N(0, I_n)\text{)} \\ &= ||(H - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)Z||^2.\end{aligned}$$

Once it is shown that $(H - \frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T)Z$ is a projection, the result follows from Lemma 8.9. □

Proof of (5). Follows from Theorem 8.10(2). □

Proof of (6). Follows by definition of F-distribution. □

With this theorem, we can now apply our intuition regarding the "closeness" of $\bar{y}$ and $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y$ and its relationship to the strength of H_0 or H_1. We construct the hypothesis test:

$$\text{Reject } H_0 \text{ when } \frac{\text{MSS}/(p-1)}{\text{RSS}/(n-p)} > F_{p-1,n-p,1-\alpha}.$$

It's important to realize here that RSS essentially remains fixed, since $\hat{y}$ is the best possible fit for y given any circumstance. Furthermore, it serves as a substitute for σ^2. If instead σ^2 is known, then we can just leverage (4) in the theorem above.

A Primer on projections

Definition A.1. A matrix $P \in R^{n\times n}$ is called a projection matrix if $P^T = P$ and $P^2 = P$.

Example A.2. Let $\mathbb{1}_n = (1, ..., 1)^T \in \mathbb{R}^n$. Then $\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T$ is a projection matrix. For some $y = (y_1, ..., y_n)$, then

$$\frac{1}{n}\mathbb{1}_n\mathbb{1}_n^T y = \left(\frac{1}{n}\sum_{i=1}^{n} y_i\right)\mathbb{1}_n = (\bar{y}, ..., \bar{y})^T.$$

This explains why $\bar{y}$ is the MLE of μ in the case of $y_1, ..., y_n \sim_{\text{i.i.d.}} N(\mu, \sigma^2)$.

Proposition A.3. *Let P be a projection matrix.*

1. *$I_n - P$ is a projection.*
2. *$P(I_n - P) = 0$ and $(I_n - P)P = 0$.*
3. *The eigenvalues of P must be either 0 or 1.*

Proof.

1. $(I_n - P)^T = I_n^T - P^T = I_n - P$ and $(I_n - P)^2 = I_n - P$, so $I_n - P$ is a projection.
2. Easy to see considering $P^2 = P$.
3. Since P is symmetric, we can write P in its eigenvalue decomposition i.e.

$$P = U\Lambda U^T, \text{ where } \Lambda \text{ is diagonal and } U^TU = UU^T = I_n.$$

The diagonal of Λ consists of the eigenvalues. Then

$$P^2 = U\Lambda U^T U\Lambda U = U\Lambda^2 U^T = U\Lambda U^T = P.$$

From this,

$$U^TU\Lambda^2U^TU = U^TU\Lambda U^TU \implies \Lambda^2 = \Lambda.$$

Hence all the eigenvalues of P must be either zero or one.

□

Theorem A.4. *Let P be a projection matrix. Suppose the eigenvalue decomposition*

$$P = U\Lambda U^T, \text{ where } \Lambda = \text{diag}\{1, ..., 1, 0, ..., 0\}.$$

Then P can be written as

$$P = \sum_{i=1}^{\text{rank}(P)} u_i u_i^T, \text{ where } u_i \text{ are eigenvectors corresponding to nonzero eigenvalues.}$$

Proof. Clear by working through the eigenvalue decomposition. □

Proposition A.5. *If P_1, P_2 are projections and $P_1P_2 = 0$ i.e. they are orthogonal projections, then $P_1 + P_2$ is a projection.*

Proof. Intuitively, this makes sense because you are effectively mapping onto a combination of orthogonal "planes". Clearly $(P_1 + P_2)^T = P_1 + P_2$, and

$$\begin{aligned}(P_1 + P_2)^2 &= P_1^2 + P_2^2 + P_1P_2 + P_2P_1 \\ &= P_1 + P_2 + 0 + (P_1P_2)^T = P_1 + P_2.\end{aligned}$$

□

Proposition A.6. *If P is a projection matrix, then* $\text{rank}(P) = \text{Tr}(P)$.

Proof. Since P is symmetric, we can write in eigenvalue decomposition i.e. $P = U\Lambda U^T$. Recalling that $\text{Tr}(BA) = \text{Tr}(AB)$, we have

$$\text{Tr}(P) = \text{Tr}(U\Lambda U^T) = \text{Tr}(\Lambda U^T U) = \text{Tr}(\Lambda).$$

□

Since Λ is just a "truncated" identity matrix per se, $\text{Tr}(\Lambda) = \text{rank}(P)$.