Mathematical Statistics 2

Koen Oostveen

April 7, 2026

Contents

1 Preliminaries

Let V be a finite-dimensional $\mathbb{R}$-vector space that is also integrable. A **random variable** X on V and a probability space (S, Σ, P) is a map $X : S \to V$. If X has a density $f : V \to \mathbb{R}$, we define the **expectation** of X through

$$EX := \int_V xf(x)\,dx$$

which is well-defined due to all densities being equal up to L^2 quotient.

Lemma 1.1. *We have that (in any basis) $(EX)^i = E(X^i)$, where $X^i : S \to \mathbb{R}$ is the i-th component of X (pointwisely).*

Lemma 1.2. *For any two random variables X, Y on the same measure and vector spaces we have $E(X + Y) = EX + EY$. That is, E is a linear operator from the space of random variables to V. The space of random variables is also a vector space (actually a module), where the field (ring) is the random variables $S \to \mathbb{R}$.*

Proof. Go into a basis and establish the equality componentwise through elementary results of probability theory. □

Lemma 1.3. *Let $A : V \to W$ be a linear map on vector spaces V, W. Then it holds that for any random variable X we have $E[AX] = AE[X]$.*

Proof. Trivial due to linearity. We can pick a basis $e_1, \ldots, e_n$ for V for example to have

$$E[AX] = A(e_i)E[X^i] = A(e_i)(EX)^i = A(EX)$$

□

Definition 1.1. The **covariance operator** $\operatorname{Cov}(X)$ for a random variable X is given in a basis through

$$\operatorname{Cov}(X)^i_j := \operatorname{Cov}(X^i, X^j)$$

Lemma 1.4. *On an inner product space with an orthonormal basis, we also have that*

$$\operatorname{Cov}(X) = E[\langle X - EX, X - EX\rangle]$$

where the inner product is understood componentwise I suppose.

Lemma 1.5. *Let $A : V \to W$ be a linear map and let X be a random variable on V. We have that*

$$\operatorname{Cov}(AX) = A \operatorname{Cov}(X) A^*$$

where A^* is the Hermitian adjoint. Remember that this is the unique linear map such that for any two vectors $x, y \in V$ we have $\langle A(x), y\rangle = \langle x, A^*(y)\rangle$.

Proof. We pick an orthonormal basis and we work it out. Remember that on an orthonormal basis we have the conjugate transpose blabla. Observe that

$$\begin{aligned}
\operatorname{Cov}(AX)^i_j &= \operatorname{Cov}(AX^i, AX^j) \\
&= \operatorname{Cov}(A^i_k X^k, A^j_k X^k) \\
&= A^i_k A^j_k \operatorname{Cov}(X^k, X^k) \\
&= A^i_l (A^*)^k_j \operatorname{Cov}(X^l, X^k) \\
&= A^i_l \operatorname{Cov}(X)^l_k (A^*)^k_j \\
&= (A \operatorname{Cov}(X))^i_k (A^*)^k_j \\
&= (A \operatorname{Cov}(X) A^*)^i_j
\end{aligned}$$

□

Definition 1.2. A linear map $A : V \to W$ on inner product spaces is said to be **positive semi-definite** is for all $v \in V$ we have $\langle v, A(v)\rangle \geq 0$.

Lemma 1.6. *The covariance operator is positive semi-definite.*

Definition 1.3. A random variable X is said to be **multivariate normally distributed** with mean $\mu \in V$ and invertible covariance matrix Σ if its density $f_X : V \to \mathbb{R}$ satisfies

$$f_X(x) = \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}\langle x-\mu, \Sigma^{-1}(x-\mu)\rangle\right)$$

and we write $X \sim N(\mu, \Sigma)$.

Lemma 1.7. *If $X \sim N(\mu, \Sigma)$ then $EX = \mu$ and $\mathrm{Var}(X) = \Sigma$.*

Remark 1.1. You don't necessarily have that if X and Y are normal, then (X, Y) is multivariate normal. Example, let $X \sim N(0, 1)$. Let ε be independent noise, any r.v, i.e. $\mathrm{Cov}(X, \varepsilon) = 0$, such that also $P(\varepsilon = 1) = P(\varepsilon = -1) = 1/2$. Let $Y = \varepsilon X$. We can show that X and Y are distributionally equal, i.e. $P(Y \leq y) = P(X \leq y)$. Then the support of the standard normal is not all of V, namely, only the subset where $Y = X$ or $Y = -X$.

Lemma 1.8. *If X and Y are normally distributed and independent, then together they are bivariate normal.*

Proof. The joint density can be computed. Let $X \sim N(\mu, \sigma^2)$ and $Y \sim N(\widetilde{\mu}, \widetilde{\sigma}^2)$. Let $\mu' := (\mu, \widetilde{\mu})$ and

$$\Sigma := \begin{bmatrix} \sigma^2 & 0 \\ 0 & \widetilde{\sigma}^2 \end{bmatrix}, \qquad \Sigma^{-1} = \begin{bmatrix} 1/\sigma^2 & 0 \\ 0 & 1/\widetilde{\sigma}^2 \end{bmatrix}$$

Then we have $(X, Y) \sim N(\mu', \Sigma)$. We can see this by

$$\begin{aligned} f_{X,Y}(x, y) &= f_X(x) f_Y(y) \\ &= \frac{1}{\sqrt{2\pi\sigma^2}\sqrt{2\pi\widetilde{\sigma}^2}} \exp\left(-\frac{1}{2}\left(\left(\frac{x-\mu}{\sigma}\right)^2 + \left(\frac{y-\widetilde{\mu}}{\widetilde{\sigma}}\right)^2\right)\right) \\ &= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}\langle (x,y) - \mu', \Sigma^{-1}((x,y)-\mu')\rangle\right) \end{aligned}$$

□

Corollary. *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be i.i.d. Then $X := (X_1, \ldots, X_n)$ is multivariate normal with $\mu' := (\mu, \ldots, \mu)$ and $\Sigma := \sigma^2 \,\mathrm{id}$ so that $X \sim N(\mu', \Sigma)$.*

Lemma 1.9. *Let V and W be appropriate vector spaces and let X be a random variable on V. Let $A : V \to W$ be linear and invertible. Let $Y := AX$. We have that*

$$f_Y(y) = f_X(A^{-1}(y))|\det(A^{-1})|$$

Proof. The density is the unique function that satisfies that for all measurable $U \subseteq W$ we have

$$P(Y \in U) = \int_U f_Y(y)\, dy$$

So we compute the above:

$$\begin{aligned} P(Y \in U) &= P(AX \in U) \\ &= P(X \in A^{-1}(U)) \\ &= \int_{A^{-1}(U)} f_X(x)\, dx \\ &= \int_U f_x(A^{-1}(y))|\det(A^{-1})|\, dy \end{aligned}$$

□

Lemma 1.10. *Let V and W be appropriate vector spaces and let $X \sim N(\mu, \Sigma)$ on V. Let $A : V \to W$ be linear and invertible. We have that*

$$AX \sim N(A\mu, A \circ \Sigma \circ A^*)$$

where A^* is the Hermitian adjoint. Recall that this is the unique map that satisfies

$$\langle A(x), y\rangle = \langle x, A^*(y)\rangle$$

Proof. We go into components in an orthonormal basis. We then compute the density. Let $Y := AX$.

$$\begin{aligned}
f_Y(y) &= f_X(A^{-1}(y))|\det(A^{-1})| \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}\langle A^{-1}(y) - \mu, \Sigma^{-1}(A^{-1}(y) - \mu)\rangle\right)|\det(A^{-1})| \\
&= \frac{1}{\sqrt{\det(2\pi\Sigma)}} \exp\left(-\frac{1}{2}\langle A^{-1}(y - A\mu), (A^* \circ (A^*)^{-1} \circ \Sigma^{-1} \circ A^{-1})(y - A\mu)\rangle\right)|\det(A)|^{-1} \\
&= \frac{1}{\sqrt{\det(2\pi A^2\Sigma)}} \exp\left(-\frac{1}{2}\langle y - A\mu, (A \circ \Sigma \circ A^*)^{-1}(y - A\mu)\rangle\right) \\
&= \frac{1}{\sqrt{\det(2\pi A^2\Sigma)}} \exp\left(-\frac{1}{2}\langle y - A\mu, (A \circ \Sigma \circ A^*)^{-1}(y - A\mu)\rangle\right)
\end{aligned}$$

We are done if we can show that $\det(A^2 \circ \Sigma) = \det(A \circ \Sigma \circ A^*)$. Well, we have that $\det(A^*) = \det(A)$ due to previous linear algebra results, and because everything is invertible this is clear. □

Lemma 1.11. *Let $X_1, \ldots, X_n \sim N(\mu, \sigma^2)$ be independent. Then*

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

where $\bar{X} := \frac{1}{n}\sum_{i=1}^n X_i$ and $\widehat{\sigma}^2 := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$.

Proof. **TODO** Let $X := (X_1, \ldots, X_n)$.

$$(n-1)\widehat{\sigma}^2 = \left\langle X - \bar{X}, X - \bar{X}\right\rangle$$

□

Theorem 1.1 (Multivariate CLT)**.** *Let $(X_k)_{k\in\mathbb{N}^*}$ be a sequence of random variables in $\mathbb{R}^d$, independent and identically distributed (iid). Suppose for all $k \in \mathbb{N}^*$ we have $EX_k = \mu$ for some $\mu \in \mathbb{R}$ and* $\operatorname{var}(X_k) = \sigma^2$ *for $\sigma > 0$. Then*

$$\sqrt{n}(\overline{X}_n - \mu) \to N(0_{\mathbb{R}^d}, \Sigma)$$

2 Regression

Let $\beta_0 \in \mathbb{R}$ and $\beta \in \mathbb{R}^n$, $\varepsilon_1, \dots, \varepsilon_m$ be independent noise with mean 0, let additionally for $i = 1, \dots, m$ we have X_i be random variables in $\mathbb{R}^n$. We say that $Y := (Y_1, \dots, Y_m)$ follows a **linear model** if

$$Y_i = \beta_0 + \beta_j X_i^j + \varepsilon_i$$

Let

$$X := \begin{bmatrix} 1 & X_1^1 & \cdots & X_1^n \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_m^1 & \cdots & X_m^n \end{bmatrix}, \beta := \begin{bmatrix} \beta_0 \\ \vdots \\ \beta_n \end{bmatrix}$$

so that we also have $Y = X\beta + \varepsilon$.

Lemma 2.1. *If $\varepsilon \sim N(0, \sigma^2 \operatorname{id})$ then $Y \sim N(X\beta, \sigma^2 \operatorname{id})$ (where X_i are not assumed to be random variables but concrete data).*

Lemma 2.2. *Let $A : \mathbb{R}^n \to \mathbb{R}^m$ be linear (where we equip $\mathbb{R}^d$ with the standard inner product). Let $b \in \mathbb{R}^m$. If $A^T A$ is invertible, then*

$$\arg\min_{x \in \mathbb{R}^n} \|Ax - b\|_2^2 = (A^T A)^{-1}(A^T b)$$

Proof. Let $y := (A^T A)^{-1}(A^T b)$. Let $r := b - Ay$. We claim that for all $p \in \operatorname{im} A$ we have $\langle r, p \rangle = 0$. Let $\widetilde{p}$ be a preimage of p, then

$$\langle r, p \rangle = \left\langle b - A(A^T A)^{-1}(A^T b), A\widetilde{p} \right\rangle = \left\langle A^T b - A^T b, \widetilde{p} \right\rangle = 0$$

Consider that

$$\begin{aligned} \|Ax - b\|_2^2 &= \|A(x - y) - (b - Ay)\| \\ &= \|A(x - y)\|_2^2 + \|b - Ay\|_2^2 \\ &\geq \|A(x - y)\|_2^2 \end{aligned}$$

where we have equality if and only if $A(x - y) = 0 \implies x = y$, so y is the unique minimizer. □

Theorem 2.1. *If $X^* \circ X$ is invertible, then the MLE and the least-squares estimator $\widehat{\beta}$ for β coincide and*

$$\widehat{\beta} = (X^* \circ X)^{-1}(X^*(Y))$$

with

$$\widehat{\beta} \sim N(\beta, \sigma^2 (X^* \circ X)^{-1})$$

Proof. The likelihood of Y is

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left(-\frac{1}{2} \left(\frac{\|x - X\beta\|}{\sigma} \right)^2 \right)$$

This is maximized by

$$\widehat{\beta} := \operatorname*{argmin}_{\beta} \|Y - X\beta\|$$

which is exactly the least-squares estimator given by solving the normal equation. It is unbiased due to

$$E\widehat{\beta} = \beta + E(X^* \circ X)^{-1}(X^*(\varepsilon)) = \beta$$

and normality and the other properties can be derived from the earlier lemma's. □

Remark 2.1. If $X^* X$ is not invertible, we can interpret that as the minimum being attained by infinitely many β, or alternatively that there are simply multiple linear models possible given the data. Such a model is called **non-identifiable**.

3 Logistic regression

Behold, a majestic smooth interpolator and actually bijection between $(0,1)$ and $\mathbb{R}$: $f : \mathbb{R} \to (0,1)$ defined by $x \mapsto f(x) := (1+\exp(-x))^{-1}$. So if we want to do a regression on a categorical variable $X \in \{A, B\}$, we can simply model the probability of category A, by taking a linear model (something we understand well) and stuffing it into the logistic function, to get a probability out. I.e. we use the model

$$P(X = A) = (1 + \exp(-(\beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n)))^{-1}$$

Now let's invert the logistic function, which is by the way invertible as we saw. For $x \in \mathbb{R}$ and $y \in (0,1)$ we have

$$y = (1+\exp(-x))^{-1} \iff x = -\log(y^{-1} - 1) = \log\left(\frac{y}{1-y}\right) =: \operatorname{logit}(y)$$

i.e. we define the **logit** function to be the inverse of the logistic function. We are left with

$$\operatorname{logit}(P(X = A)) = \beta_0 + \beta_1 x_1 + \ldots + \beta_n x_n$$

If we want to now make an estimator, we got to have a statistical model. We still observe data like before and assume that the outputs Y_i take either 0 or 1 with probability p_i, i.e. Bernoulli distributed. Let us compute the likelihood

$$L = \prod_{i=1}^{n} (1-p_i)^{1-Y_i} p_i^{Y_i}$$

so that the log-likelihood is

$$\log L = \sum_{i=1}^{n} (1-Y_i)\log(1-p_i) + Y_i \log(p_i)$$

where

$$p_i = 1/(1 + \exp(-(\beta_0 + \beta_1 x_{i,1} + \ldots + \beta_n x_{i,n})))$$

well, good luck maximizing that…

3.1 Extra: neural networks

Compare: a linear model can be seen as just a vector $\beta \in \mathbb{R}^n$ and intercept $\beta_0 \in \mathbb{R}$ such that the model can be defined by a function $f : \mathbb{R}^n \to \mathbb{R}$ such that $x \mapsto \langle x, \beta\rangle + \beta_0$, whereas a (shallow) neural network is defined as a similar function such that $\sigma : \mathbb{R}^m \to \mathbb{R}^k$ is a (continuous) activation function, $W : \mathbb{R}^n \to \mathbb{R}^m$ a linear map, and $a \in \mathbb{R}^k$ some sort of coefficients, then the model f is defined by

$$f(x) := \langle a, \sigma(Wx)\rangle$$

and of course if you choose σ, W and a appropriately, you recover the case of linear regression. Of course in general you cannot say anything about σ apart from maybe convexity or continuity conditions, but in generality you cannot guarantee much, so you get a nonconvex optimization problem out, you require some sort of gradient descent algorithm (you assume the whole thing is partially differentiable on some basis).

4 The χ^2 goodness-of-fit test

In lower generality, we have considered before tests of proportions, namely, a model with a parameter $p \in [0,1]$, and we consider for some $p_0 \in [0,1]$ the hypothesis testing problem

$$H_0 : p = p_0 \qquad \text{vs.} \quad H_1 : p \neq p_0$$

But what if we want to consider multiple proportions, namely, what if we measure proportions $p_1, \dots, p_n \in [0,1]$ and we want to test whether at least one $p_i \neq p_j$ for $i \neq j$. The setup is as follows.

We measure categorical random variables $X_1, \dots, X_n \in \{1, \dots, k\}$, we have $k \in \mathbb{N}$ categories. Let $p_1, \dots, p_k \in [0,1]$ be the proportions of occurrences of categories $1, \dots, k$ respectively, which we measure of course by (for $i = 1, \dots, k$)

$$p_i := \frac{1}{n} \sum_{j=1}^{n} \mathbf{1}\{X_j = i\}$$

If $p_1, \dots, p_n$ are the true proportions (H_0) then obviously

$$p_1 + \ldots + p_n = 1$$

or in other words $P(X_i = j) = p_j$ from which the above obviously follows. We first define the random variables

$$N_j := \sum_{j=1}^{n} \mathbf{1}\{X_j = i\} =: \#\{i : X_i = j\}$$

and then the random variable $N := (N_1, \dots, N_k) \in \mathbb{R}^k$ which then by definition follows a **multinomial distribution**, we write $N \sim MN(n, (p_1, \dots, p_k))$, where n is obviously the amount of trials. Using combinatorical arguments we can show that

$$P(N = (n_1, \dots, n_k)) = \binom{n}{n1, \dots, n_k} p_1^{n_1} \cdot \ldots \cdot p_k^{n_k} \mathbf{1}\{n_1 + \ldots + n_k = 1\}$$

where

$$\binom{n}{n_1, \dots, n_k} := \frac{n!}{n_1! \cdot \ldots \cdot n_k!}$$

We can see that this is consistent with the binomial coefficient, since for $N \sim B(n,p)$ we have

$$P(N = k) = \binom{n}{k} p^k (1-p)^{n-k} = \binom{n}{k, n-k} p_1^{n_1} p_2^{n_2}$$

where $n_1 := k$, $n_2 := n - k$, $p_1 := p$, $p_2 := 1 - p$. We can show that additionally for $N \sim MN(n, (n_1, \dots, n_k))$

$$EN = (np_1, \dots, np_k) \qquad \text{(easily consistent with binomial dist)}$$

$$\mathrm{Cov}(N)_j^i = \mathrm{Cov}(N_i, N_j) = \begin{cases} np_i(1-p_i) & \text{if } i = j \\ -np_i p_j & \text{if } i \neq j \end{cases}$$

which is once again consistent with the binomial distribution, so we have found a reasonable generalization. And just like the binomial distribution, we have asymptotic normality, namely for 'large n' we have

$$N \sim N(EN, \mathrm{Cov}(N))$$

With rescaling we can find a standard multivariate normal, namely

$$Z := \left(\frac{N_1 - np_1}{\sqrt{np_1}}, \dots, \frac{N_k - np_k}{\sqrt{np_k}} \right)$$

with $EZ = 0$ and

$$\mathrm{Cov}(Z) = \mathrm{id}(1 - p_1 - \ldots - p_k)$$

and from that it (apparently) follows that (approximately)

$$\sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j} \sim \chi^2_{k-1}$$

which we can compute quantiles on and that way do hypothesis testing. Namely, we test

$$H_0 : P(X_i = j) = p_j, \quad i = 1, \ldots, k$$

where we reject if

$$\underbrace{\sum_{j=1}^{k} \frac{(N_j - np_j)^2}{np_j}}_{=:X} > c_\alpha$$

with α the level and $P(X > c_\alpha) = \alpha$.

5 Contingency tables / The χ^2 independence test

Let's figure out if two samples of categorical variables are related in some sense. We observe a sample of n pairs of categorical random variables (X_i, Y_i) for every $i = 1, \ldots, n$, where $X_i \in \{1, \ldots, s\}$ and $Y_i \in \{1, \ldots, t\}$, with proportions $p_1, \ldots, p_s$ and $q_1, \ldots, q_t$ respectively. We define the symbol

$$r_{k,l} := P(X_i = k, Y_i = l)$$

Now if we have independence of the X and Y variables, we have

$$r_{k,l} = P(X_i = k, Y_i = l) = P(X_i = k)P(Y_i = l) = p_k q_l$$

for all k, l, which is then our null hypothesis H_0, whereas the alternative states that independence does not hold for at least one pair k, l. Let's find an estimator.

$$\widehat{r}_{k,l} = \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i = k, y_i = l\}$$

and likewise we can find estimators $\widehat{p}_k$ and $\widehat{q}_l$. We can from there show that the statistic

$$T := n \sum_{k,l} \frac{(\widehat{r}_{k,l} - \widehat{p}_k \widehat{q}_l)^2}{\widehat{p}_k \widehat{q}_l}$$

is such that $T \sim \chi^2_{(s-1)(t-1)}$ approximately / asymptotically. Yay.

6 Nonparameteric tests

Hypothesis tests that do not involve parameters of distributions. Given i.i.d. $X_1, \ldots, X_n$ random variables. We pose the testing problem: $H_0 : X_1$ generated from normal distribution v.s. not of course. We do not know a priori anything about the distribution of the X_i's. Advantage: we do not need many assumptions, disadvantage: it will be not very powerful.

6.1 Test for normality

Well, we can use the fact that $X_1, \ldots, X_n$ are normal under H_0. So, we can measure how much the data 'deviates' from the normal distribution. We use the skewness and kurtosis for this.

Definition 6.1. Let X be a random variable with finite 2nd and 3rd moment, mean μ, and standard deviation $\sigma > 0$. We define the **skewness**, $\gamma_1(X)$, of X by

$$\gamma_1(X) := E[(X - \mu)^3/\sigma^3]$$

recall that $\sigma := \sqrt{\mathrm{Var}(X)}$.

Lemma 6.1. *If the p.d.f. of X is symmetric around μ, then the skewness is zero.*

Proof. Let $f : \mathbb{R} \to [0, \infty)$ be the p.d.f. of X. Then

$$\begin{aligned}
\gamma_1(X) &= E[(X - \mu)^3/\sigma^3] \\
&= \int_{\mathbb{R}} \frac{(x-\mu)^3}{\sigma^3} f(x)\,\mathrm{d}x \\
&= \frac{1}{\sigma^3} \int_{\mathbb{R}} (x-\mu)^3 f(x)\,\mathrm{d}x \\
&= \frac{1}{\sigma^3} \underbrace{\int_{\mathbb{R}} x^3 f(x+\mu)\,\mathrm{d}x}_{0} = 0
\end{aligned}$$

where the last step is justified, because the integrand $x \mapsto x^3 f(x + \mu)$ is odd,

$$(-x)^3 f(-x + \mu) = -x^3 f(x + \mu)$$

□

Corollary. *The skewness of a normally distributed X is* 0.

Definition 6.2. Let X be a random variable with finite 2nd and 4th moment, mean μ, and standard deviation $\sigma > 0$. We define the **excess kurtosis**, $\gamma_2(X)$, of X by

$$\gamma_2(X) := E[(X - \mu)^4/\sigma^4] - 3$$

Why the -3? This is exactly because the normal distribution that way also has excess kurtosis 0.

6.1.1 Jarque-Bera test

We measure the sample skewness and kurtosis by replacing the expectations and moments by estimators. Let $\bar{X}$ be the sample mean, let

$$s_X := \sqrt{\frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^2}$$

Then we use these estimators as follows:

$$S := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^3/s_X^3, \qquad K' := \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})^4/s_X^4 - 3$$

One can show that the statistic

$$JB := \frac{n}{6}\left(S^2 + K'^2/4\right)$$

has approximately / asymptotically a χ^2_2 distribution. So we can do hypothesis testing! (for large samples). Though, the power of this test is quite low, especially for distributions that have similar 3rd and 4th moments like the normal distribution, so that they have skewness and excess kurtosis of 0.

Clearly, just using certain statistics, namely $\bar{X}$, s_X, skewness, kurtosis, etc. tell you information about the distribution, but there will always be an information loss. But if we use the order statistics, no information will be lost, you know **every** quantile of the measured distribution as good as you can get it with your sample. In fact, we can use the expectation of the order statistics given a certain sample size n to make a **QQ-plot**. The follow test simulates this plot.

6.1.2 Shapiro-Francia test

Let $X_1, \ldots, X_n \sim N(0,1)$ be an i.i.d. sample of the standard normal distribution. Let

$$X_{(1)} \leq \ldots \leq X_{(n)}$$

be the corresponding order statistics. Let

$$X' := (X_{(1)}, \ldots, X_{(n)}), \qquad m := (EX_{(1)}, \ldots, EX_{(n)})$$

The **Shapiro-Francia test** uses the test statistic $W' := \widehat{\text{Corr}}$, i.e. $W'(X', m) := \widehat{\text{Corr}}(X', m)$. In practice this means we compute

$$W' = \frac{\sum_{i=1}^n (X_{(i)} - \bar{X})(m_i - \bar{m})}{\sqrt{\sum_{i=1}^n (X_{(i)} - \bar{X})^2 \sum_{i=1}^n (m_i - \bar{m})^2}}$$

The test clearly works like this: we reject if $W' < c$ for appropriate critical value c.

6.1.3 Shapiro-Wilk test

Whereas the previous section essentially used EX', this test also uses the 'second moment', in the sense of the covariance matrix $\text{Cov}(X') =: V$, which is assumed to be invertible / positive definite. Let $a := m^T V^{-1} / \left\| m^T V^{-1} \right\|_2$.

Remark 6.1. Recall that for linear $A : \mathbb{R}^n \to \mathbb{R}^m$ where both spaces are equipped with the standard inner product and a vector $w \in \mathbb{R}^m$, that $w^T A := (A^T w)^T := (A^* w)^*$, where the transpose is essentially the index-pulling method that we get from choosing a basis, the standard basis. This vector satisfies that for $v \in \mathbb{R}^n$, $(w^T A)v = w^T(Av)$, i.e. it guarantees something about an inner product, namely that $\langle w^T A, v\rangle = \langle w, Av\rangle = \langle A^* w, v\rangle$. Aha!

Anyway, with such a, we get that $\|a\|_2 = 1$, encoding 'weighted averages' of covariances scaled by the mean. We can use this to compute the standard deviation, in the sense that

$$\langle a, X\rangle = \langle m, V^{-1}X\rangle / \left\| m^T V^{-1} \right\|_2 = \sigma$$

So we define the statistic $W := \frac{\langle a, X\rangle}{\widehat{\sigma}}$, where

$$\widehat{\sigma} := \frac{1}{n-1}\sum_{i=1}^n (X_i - \bar{X})^2$$

Now we can show that $W \leq 1$, and $EW = 1$ under the null hypothesis, so we reject if $W < c$ for some appropriate quantile c, that is usually found through simulations.

Remark 6.2.

- The test can encounter problems if many values are equal.

- The test rarely rejects H_0 when the sample size is small.
- It rejects H_0 for tiny differences from normality when n is large.
- A major problem is that normality is the null hypothesis.
- In practice we are rather interested in the opposite problem.
- For the construction it is crucial that data are normal under H_0 .
- The limiting distribution (after standardization) is not χ^2, but a weighted infinite sum of shifted squared standard normals. Critical regions are found by Monte-Carlo simulations.

In practice, the Shapiro-Wilk test is used together with other measures for normality, such as kurtosis, histograms and QQ-plots.

6.2 Permutation test (for distribution)

Let $X_1, \dots, X_n$ and $Y_1, \dots, Y_m$ be i.i.d. with means μ_X, μ_Y. We define the testing problem $H_0 : \mu_X = \mu_Y$ v.s. the negation. A simple idea for a test statistic is just $T := \bar{X} - \bar{Y}$ which has mean 0, so we reject if $|T| > c$ for a critical value $c > 0$. But because we have not assumed anything about the distribution, this critical value (for the desired level) cannot be found. Well, if we can somehow approximate the c.d.f $P(T \leq t)$, we can find the quantiles.

We can approximate this, however, by using the data that we already have. Additionally, under H_0, X and Y have identical distributions. More generally, we can have a testing problem with distributions P, Q such that: $X_1, \dots, X_n \sim P$, $Y_1, \dots, Y_m \sim Q$, and we test

$$H_0 : P = Q \quad \text{v.s.} \quad H_1 : P \neq Q$$

Definition 6.3. Let $n \in N^*$. A **permutation** of **arity** n is a bijection $\phi : \{1, \dots, n\} \to \{1, \dots, n\}$.

The idea is that under H_0, permuting the data should not affect T much if they are sampled from the same data, as $\bar{X} = \bar{Y}$ under H_0. So we can use **every** permutation (of which there exists finitely many, namely $(m+n)!$) of the vector $Z := (X_1, \dots, X_n, Y_1, \dots, Y_m)$, and find the p-value by testing how often a permutation goes past the mean in a certain direction (which under H_0 should be about half the time). We define $\widehat{p}$, an estimator for the p-value, by first labelling $\phi_1, \dots, \phi_N$ all permutations of $m+n$ elements, and

$$\widehat{p}(Z) := \frac{1}{N} \sum_{i=1}^{n} \mathbf{1}\{T(\phi_i(Z)) \geq T(Z)\}$$

The test clearly becomes $\phi(Z) := \mathbf{1}\{\widehat{p}(Z) < \alpha\}$ at level α. The Type I error, or empirical level, is

$$P_0(\phi = 1) = P_0(\widehat{p}(Z) < \alpha) \leq \alpha$$

How can we see the final step? Well, under H_0, we note that for any i,

$$P(T(\phi_i(Z)) \geq T(Z)) = P(T(\phi_i(Z)) \leq T(Z)) = 1/2$$

i.e.

$$\sum_{i=1}^{n} \mathbf{1}\{T(\phi_i(Z)) \geq T(Z)\} \sim B(N, 1/2)$$

6.3 Sign test (for median)

Definition 6.4. Let X be a random variable. We say that $m \in \mathbb{R}$ is the **median** of X if

$$P(X < m) \leq 1/2 \wedge P(X > m) \leq 1/2$$

If it is unique, we write $\text{Med}(X) := m$. If X has quantile function Q_X, $Q(1/2)$ is clearly a median.

Lemma 6.2. *Let $X \sim N(\mu, \sigma^2)$, then* $\text{Med}(X) = \mu$.

Proof. Consider that $P(X < \mu) = P(X > \mu) = 1/2$ so it is clearly a median. Now we show that is unique. Let $\varepsilon \neq 0$ be a deviation to the median, such that $\widetilde{m} := m + \varepsilon$ is an alternative median. Then

$$\begin{aligned} P(X < \widetilde{m}) &= P(m < X < \widetilde{m}) + P(X < m) \\ &= \int_m^{\widetilde{m}} f_X(x)\,dx + 1/2 \\ &\geq (\widetilde{m} - m) \sup_{x \in [m, \widetilde{m}]} f_X(x) + 1/2 \\ &> 1/2 \end{aligned}$$

so $\widetilde{m}$ is not a median, so by contradiction we are done. □

Definition 6.5 (Sign test)**.** Let $X_1, \ldots, X_n$ be iid that have a unique median, let $m_0 \in \mathbb{R}$. We test $H_0 : \text{Med}(X_1) = m_0$ v.s. $H_1 : \text{Med}(X_1) \neq m_0$. Define the test statistic

$$T := \sum_{i=1}^n \mathbf{1}\{X_i \leq m_0\}$$

Then we reject if $|T| > c$ for certain critical value c. This is called the **sign test** since

$$T = \sum_{i=1}^n \mathbf{1}\{\text{sign}(X_i - m_0) = -1\}$$

Lemma 6.3. *In this sign test, $T \sim B(n, 1/2)$ under H_0.*

Proof. Clearly, T is the sum of Bernoulli random variables $B_1, \ldots, B_n$, where under H_0 for each i

$$P(B_i = 1) = P(X_i \leq m_0) = 1/2$$

So then $T \sim B(n, 1/2)$. □

Remark 6.3. For large n one can use that $B(n, 1/2) \approx N(n/2, n/4)$.

6.4 Runs test (for distribution / stochastic ordering)

Definition 6.6. Let X, Y be random variables. We say that $X \leq Y$, or **X is stochastically less than Y**, if for all $t \in \mathbb{R}$ we have

$$F_X(t) = P(X \leq t) \geq P(Y \leq t) = F_Y(t)$$

i.e. the cdf of X is bounded below by Y. This makes $\leq$ a partial ordering.

Theorem 6.1. *Let X and Y be random variables with density f_X and f_Y, respectively. tfae*

- $X \leq Y$
- *For all increasing functions $\phi : \mathbb{R} \to \mathbb{R}$, $E[\phi(X)] \leq E[\phi(Y)]$ (so in particular $EX \leq EY$)*
- *For all strictly increasing functions $\phi : \mathbb{R} \to \mathbb{R}$, $\phi(X) \leq \phi(Y)$*
- *For all $A \subseteq \mathbb{R}$ with increasing indicator, $P(X \in A) \leq P(Y \in A)$*

Proof. Suppose $X \leq Y$. Let $\phi : \mathbb{R} \to \mathbb{R}$ be increasing. That is, for all $x, y \in \mathbb{R}$ such that $x \leq y$, we have $\phi(x) \leq \phi(y)$. Notice that for all nonnegative random variables Z

$$\begin{aligned} EZ &= \int_{\mathbb{R}} z f_Z(z)\,dz \\ &= \int_{\mathbb{R}} \int_0^\infty \mathbf{1}\{z > t\} f_Z(z)\,dt\,dz \\ &= \int_0^\infty \int_{\mathbb{R}} \mathbf{1}\{z > t\} f_Z(z)\,dz\,dt \\ &= \int_0^\infty P(Z > t)\,dt \end{aligned}$$

Therefore, since $P(\phi(X) > t) \leq P(\phi(Y) > t)$, we have the desired result. Now suppose we have indeed $E[\phi(X)] \leq E[\phi(Y)]$ for all increasing functions. Clearly, if ϕ is strictly increasing it is in particular increasing. Let $t \in \mathbb{R}$, we want to show that $P(\phi(X) > t) \leq P(\phi(Y) > t)$. Let $g_n : \mathbb{R} \to \mathbb{R}$ be an increasing step function that converges to $g : \mathbb{R} \to \mathbb{R}$ defined by $x \mapsto \mathbf{1}\{u > t\}$. Then by the assumption we have that $E[(g_n \circ \phi)(X)] \leq E[(g_n \circ \phi)(Y)]$. Then

$$P(\phi(X) > t) = \lim_{n\to\infty} E[(g_n \circ \phi)(X)] \leq \lim_{n\to\infty} E[(g_n \circ \phi)(Y)] = P(\phi(Y) > t)$$

So we are done for this part. The next implications are trivial. □

Definition 6.7 (Rank of sample)**.** Let $\pi : \{1, \ldots, n\} \to \{1, \ldots, n\}$ be a permutation (bijection) that orders a sample $z_1, \ldots, z_n$ into $z_{\phi(1)}, \ldots, z_{\phi(n)}$, so that the latter is sorted w.r.t $\leq$. We say that the **rank** of z_i for all i is $\phi(i)$.

Definition 6.8 (Runs test)**.** Let $X_1, \ldots, X_n \in \mathbb{R}$, $Y_1, \ldots, Y_m \in \mathbb{R}$ both from an iid sample (suppose $X \sim F$ and $Y \sim G$). We test $H_0 : F = G$ versus $H_1 : F \neq G$ by rejecting H_0 if $R < c$ for appropriate critical value c, where R is the **amount of runs**, given by the following. We take the sample $X_1, \ldots, X_n, Y_1, \ldots, Y_m$ and find the ranks. Then we color the sequence $\pi(1), \ldots, \pi(n+m)$ by consecutive sequences, and count how many 'runs' we have, this is how we define R.

Theorem 6.2. *Under H_0, we have that*

$$P(R = r) = hugemonsterhelp$$

Remark 6.4. R is asymptotically normal! So for large sample sizes, we can use that.

Remark 6.5. R is a measure of stochastic ordering, so an alternative reformulation of

$$H_1 : \text{distributions are stochastically ordered}$$

6.5 Wilcoxon rank sum test (for distribution)

Definition 6.9 (Rank sum test)**.** Let $X_1, \ldots, X_n \sim F$ and $Y_1, \ldots, Y_m \sim G$ be independent with continuous density, we test similarly again $H_0 : F = G$ v.s. $H_1 : F \neq G$. Let

$$W := \sum_{i=1}^{n} R(X_i)$$

Lemma 6.4. *Clearly using well-known sum identities and bounding $R(\cdot)$*

$$n(n+1)/2 \leq W \leq (n+m)(n+m+1)/2 - n(n+1)/2$$

Recall that there are $(n+m)!$ permutations of arity $n+m$. Every outcome of

$$(R(X_1), \ldots, R(X_n), R(Y_1), \ldots, R(Y_m))$$

is equally likely, so $EW = n(n+m+1)/2$. Let $R_i := R(X_i)$. We can show this by observing that R_i is uniform among $\{1, \ldots, n+m\}$, so $E[R_i] = \frac{n+m+1}{2}$. Therefore $EW = n(n+m+1)/2$. We can also show that

$$\mathrm{Var}(R_i) = \frac{(n+m)^2 - 1}{12}, \qquad \mathrm{Cov}(R_i, R_j) = -\frac{n+m+1}{12}$$

$$\begin{aligned}\mathrm{Var}(W) &= \mathrm{Var}\left(\sum_{i=1}^{n} R_i\right) \\ &= \sum_{i=1}^{n}\sum_{j=1}^{n} \mathrm{Cov}(R_i, R_j) \\ &= \sum_{i=1}^{n} \frac{(n+m)^2-1}{12} - (n-1)\frac{n+m+1}{12} \\ &= \sum_{i=1}^{n} \frac{n^2+2nm+m^2-1}{12} - \frac{n^2+nm+n-n-m-1}{12} \\ &= \sum_{i=1}^{n} m\frac{n+m+1}{12} = nm\frac{n+m+1}{12}\end{aligned}$$

Using this data we can use asymptotic normality for large samples and the CLT.

Theorem 6.3. *Under H_0, W has an approximate distribution of*

$$N\left(\frac{n(n+m+1)}{2}, \frac{nm(n+m+1)}{12}\right)$$

6.6 Wilcoxon signed rank sum test (for median)

Suppose that X is a continuous random variable **with symmetric pdf around the mean**. We want to test $H_0 : m = m_0$ for some candidate median $m_0 \in \mathbb{R}$ versus $H_1 : m \neq m_0$ or similar. We define the **signed rank sum statistic** W by

$$W := \sum_{i=1}^{n} \mathbf{1}\{X_i \geq m_0\} R_i$$

where the ranks $R_1, \ldots, R_n$ of the sample $X_1, \ldots, X_n$ are ordered along $|X_i - m_0|$, i.e. the distance to the proposed median defines the ranks. So in the end, we sum those ranks of the sample that exceed the median. Under H_0, we can expect $X_i \geq m_0$ to be true approximately half of the time, and the ranks are uniformly distributed among the sample, so like before under H_0 we can expect W to hover around a certain number, we reject if $|W| > c$ (or something else depending on H_1) for an appropriate critical value c, or we compute the p-value instead with simulations.

Like before, we can compute means and variances, they come out to be

$$EW = \frac{n(n+1)}{4}, \qquad \mathrm{Var}(W) = \frac{n(n+1)(2n+1)}{24}$$

6.6.1 Ties

What if a tie occurs when computing ranks? We can ignore the problem in two ways:

- Remove the entries from the sample (especially useful when testing $H_0 : m = 0$ v.s. $H_1 : m > 0$ and the ties are zero's)
- Find all the ranks that would be distributed amongst the ties, e.g. $R_1, \ldots, R_i, \ldots, R_j, \ldots, R_n$ are the ranks of the sample, and $R_i, \ldots, R_j$ give ties, then find $\widetilde{R} := (R_i + \ldots + R_j)/(j - i)$ (the arithmetic mean) and assign that rank to every tied entry.

6.7 Summary

Remark 6.6 (Permutation test). A permutation test finds whether two samples are of the same distribution.
Advantages:

- The test is exact: we can compute the quantiles under H_0 analytically, the test statistic is Bernoulli distributed
- Can be used for lots of test statistics instead of just $T := \bar{X} - \bar{Y}$
- Because of the above we can use many hypotheses

Disadvantages:

- Computationally expensive to compute the quantiles of $B(n, 1/2)$ for large n, as well as computing permutations (for large n, the amount of permutations grows superexponentially)
- Assumes that the distribution is invariant under permutation

Remark 6.7 (Sign test). A sign test can check if the median significantly differs from a proposed median.
Advantages:

- Easy to compute
- Works for any data for which a sign can be computed (not necessarily numerical)

Disadvantage:

- Low power
- Something about paired samples?

Remark 6.8 (Runs test). A runs test finds whether two samples are of the same distribution.
Adv. **TODO**

7 The bootstrap

Let $\widehat{\theta}$ be an estimator of $\theta \in \Theta \subseteq \mathbb{R}$. We would like to quantify the uncertainty of $\widehat{\theta}$ given:

- Realizations $(x_1, \ldots, x_n)$ of an iid sample $X_1, \ldots, X_n$ of random variables
- NO access to the distribution of $\widehat{\theta}(X_1, \ldots, X_n)$ as an r.v.
- No access to any other data
- No parametric assumptions

Definition 7.1. Given realizations $x_1, \ldots, x_n$ of an iid sample $X_1, \ldots, X_n$ and an estimator $\widehat{\theta}$, we say that a **bootstrap sample** of $x_1, \ldots, x_n$ is an ordered tuple $x_1^*, \ldots, x_n^*$ of which each entry is contained in $\{x_1, \ldots, x_n\}$. I.e. it is a draw **with replacement** of n items from $x_1, \ldots, x_n$.
A **bootstrap estimate** $\widehat{\theta}^* \in \mathbb{R}$ is then the realization of $\widehat{\theta}$ under $x_1^*, \ldots, x_n^*$, i.e.

$$\widehat{\theta}^* = \widehat{\theta}(x_1^*, \ldots, x_n^*)$$

Definition 7.2 (Empirical distribution function)**.** Let $X = (x_1, \ldots, x_n) \in \mathbb{R}^n$ be some sort of sample. We say that its associated **empirical distribution function** (EDF) is a function $\widehat{F}_X(x) : \mathbb{R} \to [0, 1]$ (a type of cdf) such that

$$\widehat{F}_X(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{x_i \leq x\}$$

Remark 7.1. We can also define a **family of random variables**, namely $\widehat{F}_X(x)$ for each $x \in \mathbb{R}$ given an i.i.d. random sample $X_1, \ldots, X_n$ such that

$$\widehat{F}_X(x) := \frac{1}{n} \sum_{i=1}^{n} \mathbf{1}\{X_i \leq x\}$$

almost the same thing, conflated very commonly with the EDF, the difference is that this is a family of random variables and the other simply a function $\mathbb{R} \to [0, 1]$.

Remark 7.2. In a measure theoretic sense, $\widehat{F}_X$ is also a random variable, i.e. an estimator for F, the true cdf of $X_1, \ldots, X_n$. Notice that

$$E\widehat{F}_X(x) = \frac{1}{n} \sum_{i=1}^{n} P(X_i \leq x) = \frac{1}{n} nF(X_i \leq x) = F(X_i \leq x)$$

so in a sense, $E\widehat{F}_X = F$, so $\widehat{F}$ is unbiased for F. woah. And probably with the appropriate assumptions on F we get that it is also consistent.

Remark 7.3. Given $\widehat{\theta}_1^*, \ldots \widehat{\theta}_B^*$, B bootstrap estimates for θ, we can now use the EDF to reason about the distribution of θ empirically. The **bootstrap assumption** is that the EDF of the bootstrap estimates is a good approximation of the EDF of $X_1, \ldots, X_n$.

Definition 7.3. Let $\widehat{\theta}_1^*, \ldots \widehat{\theta}_B^*$, be bootstrap estimates for θ. We define the **standard bootstrap error**, $\widehat{\mathrm{SE}_{\mathrm{boot}}} > 0$, through

$$\widehat{\mathrm{SE}_{\mathrm{boot}}}^{\,2} := \frac{1}{B-1} \sum_{i=1}^{B} \left(\widehat{\theta}_i^* - \overline{\widehat{\theta}^*}\right)^2$$

i.e. it is like $\widehat{\sigma}^2$ but on bootstrap estimates.

Remark 7.4. The bootstrap standard error is a measure for the **variance** of $\widehat{\theta}$, even though we had absolutely no idea about its underlying distribution before.

7.1 Parametric bootstrap

If we know the underlying model, we can also get a more nuanced bootstrap. Instead of sampling from measurements $X_1, \dots, X_n \sim P_\theta$ where $\theta \in \Theta \subseteq \mathbb{R}^d$ is a parameter, we can sample $P_{\widehat{\theta}}$ (if possible).

Example 7.1. Let $X_1, \dots, X_n \sim U[0, \theta]$. Let $\theta > 0$. The MLE is $\widehat{\theta} := \max\{X_1, \dots, X_n\}$. Suppose $\widehat{\theta}$ has a realization, then we sample $X_1^*, \dots, X_n^* \sim U[0, \widehat{\theta}]$.

7.2 Confidence intervals

A lot of quantities that we want to compute can be written as functionals T such that for all probability measures P, $T(P) \in \mathbb{R}$ is the measured quantity. For example, we can write

$$\widehat{\mu}(P) := \int_\Omega x dP, \qquad \widehat{\sigma}^2(P) := \int_\Omega (x - \mu)^2 dP$$

Let $\widehat{P}$ be the probability measure derived from the EDF of measurements $x_1, \dots, x_n$. Then we can estimate $T(P)$ using $T(\widehat{P})$ clearly. How can we now construct confidence intervals for $T(P)$ using the estimator $T(\widehat{P})$? Well, just like before, we can bootstrap $\widehat{P}$ by sampling it in some sort of way, and getting a bootstrap sample $x_1^*, \dots, x_n^*$, of which the EDF induced another probability measure P^*.
The **bootstrap principle** then assumes that we can estimate the distribution of $T(\widehat{P}) - T(P)$ (the error of $T(\widehat{P})$) using $T(P^*) - T(\widehat{P})$.

Remark 7.5. In order to construct a one-sided $1 - \alpha$ confidence interval for $T(P)$, we could find c s.t.

$$P(T(\widehat{P}) - T(P) \geq c) = 1 - \alpha$$

because then we have the $(1 - \alpha)$-CI $(-\infty, T(\widehat{P}) + c)$ for $T(P)$. Using the bootstrap principle, we find this c such that

$$P(T(P^*) - T(\widehat{P}) \geq c) = 1 - \alpha$$

And similarly for other intervals. Just pretend that P^* is the actual probability distribution and good things will happen.

Remark 7.6. In this framework with functionals T, we can similarly find a 'bootstrap variance' or 'bootstrap standard deviation' like so: if we have **bootstrap replicates** $T(x^{*1}), \dots, T(x^{*B})$, we define

$$\widehat{\mathrm{Var}}^*(T(x)) := \frac{1}{B-1} \sum_{i=1}^{B} \left(T(x^{*i}) - \frac{1}{B} \sum_{j=1}^{B} T(x^{*j}) \right)^2, \quad \widehat{\mathrm{SD}}^*(T(x)) := \sqrt{\widehat{\mathrm{Var}}^*(T(x))}$$

Example 7.2. We apply the bootstrap to the median. We measure $X_1, \dots, X_n$, generate B bootstrap estimates M_n^{*i} for M_n, where $i = 1, \dots, B$, and like in the above we can compute the bootstrap variance:

$$\widehat{\mathrm{Var}}^*(M_n) = \frac{1}{B-1} \sum_{i=1}^{B} \left(M_n^{*i} - \frac{1}{B} \sum_{j=1}^{B} M_n^{*j} \right)^2$$

And what about a confidence interval? We suppose that due to the CLT and large enough sample size we can use that

$$M_n^* \sim N(M_n, \widehat{\mathrm{Var}}^*(M_n))$$

By doing so, we can simply use z-quantiles.

Definition 7.4. Let $[a, b] \subseteq \mathbb{R}$ be a $1 - \alpha$ confidence interval for a quantity $T(P)$. We say that the **converage probability** of the confidence interval is the probability

$$P(T(P) \in [a, b])$$

Remark 7.7. More generally, let $T(P_n)$ be a quantity of interest. We estimate using observed data $T(\widehat{P}_n)$. If the central limit theorem holds, and $T(\widehat{P}_n)$ is unbiased, (which can be different criteria), we obviously have that

$$\frac{T(\widehat{P}_n) - T(P_n)}{\sqrt{\mathrm{Var}(T(\widehat{P}_n))}} \to N(0,1)$$

of which we can easily compute quantiles. The quantity

$$\mathrm{Var}(T(\widehat{P}_n)) \approx \mathrm{Var}(T(P_n^*))$$

using the bootstrap (assuming consistency). Under the assumption that $T(P_n^*)$ is normally distributed with mean $T(\widehat{P})$, we can compute approximate confidence intervals

$$\begin{aligned}
1-\alpha &\approx P\left(\frac{T(\widehat{P}_n) - T(P_n)}{\sqrt{\mathrm{Var}(T(\widehat{P}_n))}} \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\right) \\
&\approx P\left(\frac{T(\widehat{P}_n) - T(P_n)}{\underbrace{\sqrt{\mathrm{Var}(T(P_n^*))}}_{=:\sigma_n^*}} \in [-z_{1-\alpha/2}, z_{1-\alpha/2}]\right) \\
&= P\Big(T(\widehat{P}) - z_{1-\alpha/2}\sigma_n^* \le T(P) \le T(\widehat{P}) + z_{1-\alpha/2}\sigma_n^*\Big)
\end{aligned}$$

Definition 7.5. The **normal bootstrap interval** for an estimator $T(\widehat{P}_n)$ that is **asymptotically normal** given bootstrap samples

$$\sigma_n^* = \sqrt{\mathrm{Var}(T(P_n^*))}$$

is the $1-\alpha$ approximate confidence interval (for $T(p_n)$)

$$[T(\widehat{P}) - z_{1-\alpha/2}\sigma_n^*, T(\widehat{P}) + z_{1-\alpha/2}\sigma_n^*]$$

Remark 7.8. Recall that for a certain $X \sim F$ where F has quantile function $Q : (0,1) \to \mathbb{R}$ satisfies (with $\alpha < \beta$)

$$P(Q(\alpha) \le X \le Q(\beta)) = P(X \le Q(\beta)) - P(X \le Q(\alpha)) = \beta - \alpha$$

So that in particular

$$P(Q(\alpha/2) \le X \le Q(1-\alpha/2)) = 1-\alpha$$

which can be used to make confidence intervals. Note that if $EX = 0$ and the pdf is symmetric, then it holds that $Q(\alpha/2) = -Q(1-\alpha/2)$, but not in general.

Definition 7.6. The **basic pivot bootstrap interval** for an estimator $T(\widehat{P}_n)$ such that

$$T(\widehat{P}_n) - T(P_n) \approx T(P_n^*) - T(\widehat{P}_n) \qquad \text{(asymptotically)}$$

is constructed by taking the quantile function Q of $T(P_n^*) - T(\widehat{P}_n)$, and then taking the $1-\alpha$ approximate confidence interval (for $T(P_n)$)

$$[T(\widehat{P}_n) - Q(1-\alpha/2), T(\widehat{P}_n) - Q(\alpha/2)]$$

Definition 7.7. The **percentile bootstrap interval** for an estimator $T(\widehat{P}_n)$ such that

$$T(P_n^*) \approx T(\widehat{P}_n) \qquad \text{(asymptotically)}$$

is constructed by taking the quantile function Q of $T(P_n^*)$, and then taking the $1-\alpha$ approximate confidence interval (for $T(P_n)$)

$$[Q(\alpha/2), Q(1-\alpha/2)]$$

Remark 7.9. These intervals are cool and all, but do they achieve good coverage? This is hard to compute, and generally speaking we can only talk about asymptotic behaviour, but we can do simulations. It turns out that the coverage is usually not very close to the desired $1-\alpha$ expected probability. This is because the above intervals may not come from a distribution that is asymptotically close to the real distribution. There is simply a bias that occurs for lower sample sizes, in the form of variance. 'Dividing out' the variance would solve this problem, which leads to studentized bootstrap intervals.

Definition 7.8. The **studentized bootstrap interval** for an estimator $T(\widehat{P}_n)$ of $T(P_n)$ given a standard deviation operator

$$\sigma(P) := \sqrt{\mathrm{Var}(T(P))}$$

and a quantile function $Q : (0,1) \to \mathbb{R}$ of the quantity

$$\frac{T(P_n^*) - T(\widehat{P})}{\sigma(P_n^*)}$$

is the $1-\alpha$ approximate confidence interval (for $T(p_n)$)

$$[T(\widehat{P}) - Q(1-\alpha/2)\sigma(\widehat{P}_n), T(\widehat{P}) - Q(\alpha/2)\sigma(\widehat{P}_n)]$$

Remark 7.10. The above follows from

$$\begin{aligned}
1-\alpha &\approx P\left(Q(\alpha/2) \le \frac{T(P_n^*) - T(\widehat{P})}{\sigma(P_n^*)} \le Q(1-\alpha/2)\right) \\
&\approx P\left(Q(\alpha/2) \le \frac{T(\widehat{P}_n) - T(P)}{\sigma(\widehat{P}_n)} \le Q(1-\alpha/2)\right) \\
&= P\Big(T(\widehat{P}_n) - Q(\alpha/2)\sigma(\widehat{P}_n) \ge T(P) \ge T(\widehat{P}_n) - Q(1-\alpha/2)\sigma(\widehat{P}_n)\Big)
\end{aligned}$$

7.3 Bootstrap failure

Example 7.3. Let $X_1, \dots, X_n \sim U[0,1]$ be i.i.d, let $T_n := \min\{X_1, \dots, X_n\}$. Then (in distribution)

$$nT_n \to \mathrm{Exp}(1)$$

Proof. We need to check that $F_n \to F$ pointwisely, where F_n is the cdf of nT_n, and F is the cdf of a Exp(1) variable, the latter is given by

$$F(x) = 1 - e^{-x}$$

We compute

$$\begin{aligned}
F_n(x) &= P(nT_n \le x) \\
&= P\Big(T_n \le \frac{x}{n}\Big) \\
&= 1 - P\Big(T_n > \frac{x}{n}\Big) \\
&= 1 - \prod_{i=1}^{n} P\Big(X_i > \frac{x}{n}\Big) \\
&= 1 - \prod_{i=1}^{n} \Big(1 - P\Big(X_i \le \frac{x}{n}\Big)\Big) \\
&= 1 - \prod_{i=1}^{n} \Big(1 - \frac{x}{n}\Big) \\
&= 1 - \Big(1 - \frac{x}{n}\Big)^n \to 1 - e^{-x}
\end{aligned}$$

There is more subtlety, though, because the c.d.f is actually given by

$$P\Big(X_i \leq \frac{x}{n}\Big) = \begin{cases} 0 & \text{if } \frac{x}{n} < 0 \\ \frac{x}{n} & \text{if } 0 \leq \frac{x}{n} \leq 1 \\ 1 & \text{if } \frac{x}{n} > 1 \end{cases}$$

So we restrict to $x > 0$ and then for large enough n we have $\frac{x}{n} < 1$, and in the case that $x < 0$ the cdfs trivially coincide. □

The result is that for appropriate n, nT_n is approximately Exp(1), and therefore

$$1 - e^{-nx} \approx P(nT_n \leq nx) = P(T_n \leq x)$$

and therefore T_n is approximately Exp(n).
Now suppose one bootstrap sample is $X_1^*, \ldots, X_n^*$, and we compute $T_n^* = \min\{X_1^*, \ldots, X_n^*\}$. We get that for all $i = 1, \ldots, n$

$$P(X_i^* = M_n) = \frac{1}{n}$$

because we are drawing with replacement, assuming the original data $X_1, \ldots, X_n$ contains no duplicates (which happens wp 1). Therefore

$$P(X_1^* \neq M_n, \ldots, X_n^* \neq M_n) = P(X_1^* \neq M_n)^n = \left(1 - \frac{1}{n}\right)^n \to e^{-1}$$

whereby

$$P(M_n^* = M_n) \to 1 - e^{-1} \neq 0$$

7.4 Hypothesis testing

Let's say we want to test

$$H_0 : \theta \in \Theta_0 \qquad \text{v.s.} \qquad H_1 : \theta \notin \theta_0$$

with a test statistic $T : \mathbb{R}^n \to \mathbb{R}$, and our rejection region is $T > c_\alpha$ for appropriate critical value c_α, such that

$$\forall \theta \in \Theta_0 : P_\theta(T(X) > c_\alpha) \leq \alpha \iff \sup_{\theta \in \Theta_0} P_\theta(T(X) > c_\alpha) \leq \alpha$$

Remark 7.11 (p-value). By definition, the p-value is the smallest significance level for which we would still reject. So for measured data $x \in \mathbb{R}^n$

$$p(x) = \inf\{\alpha \in (0,1) \mid T(x) > c_\alpha\}$$

so that

$$T(x) > c_\alpha \iff p(x) \leq \alpha$$

But also by monotonicity and the decreasing property of the test

$$T(x) > c_\alpha \iff \sup_{\theta \in \Theta_0} P_\theta(T(X) > T(x)) < \sup_{\theta \in \Theta_0} P_\theta(T(X) > c_\alpha) \leq \alpha$$

so that

$$p(x) = \inf\{\alpha \in (0,1) \mid \sup_{\theta \in \Theta_0} P_\theta(T(X) > T(x)) < \alpha\} = \sup_{\theta \in \Theta_0} P_\theta(T(X) > T(x))$$

Intuituively, this means that the **smallest level for which we would still reject** is the **upper bound of the probability of measuring** $T(x)$ or higher.

Continuing, we find an estimator $\widehat{\theta}$ and use the approximate distribution $P_{\widehat{\theta}} =: \widehat{P}_\theta$. We generate i.i.d. bootstrap samples $(X^{*1}, \dots, X^{*B}) \in (\mathbb{R}^n)^B$. We define the estimator of the p-value:

$$\widehat{p} := \widehat{P}(T(X) \geq T(x)) = \frac{1}{B} \sum_{i=1}^{B} \mathbf{1}\{T(X^{*i}) \geq T(x)\}$$

An alternative:

$$\widehat{p} := \frac{1 + \sum_{i=1}^{B} \mathbf{1}\{T(X^{*i}) \geq T(x)\}}{B+1}$$

which behaves asymptotically identical but slightly overestimates the p-value for small values of B, to not reject too often.

7.5 Asymptotic behaviour

Theorem 7.1. *Let $(X_n)_{n\in\mathbb{N}^*}$ be a sequence of i.i.d. random variables in $\mathbb{R}^d$. Let $\mu := EX_1$ and $\Sigma := \mathrm{Cov}(X_1)$. Let for every $n \in \mathbb{N}*$, there be a iid bootstrap sample of $X_1, \dots, X_n$, namely for empirical distribution $\widehat{P}_n$,*

$$X^*_{n,1}, \dots, X^*_{n,n} \sim \widehat{P}_n$$

(draws with replacement). Define additionally the sequences

$$\bar{X}_n := \frac{1}{n}\sum_{i=1}^{n} X_i, \qquad \bar{X}^*_n := \frac{1}{n}\sum_{i=1}^{n} X^*_{n,i}$$

Then the conditional distribution of $\sqrt{n}(\bar{X}^*_n - \bar{X}_n)$ on $X_1, \dots, X_n$ converges weakly to $N(0, \Sigma)$.

Remark 7.12. This can be shown using the multivariate CLT and additionally using that the bootstrap samples are random variables in a sense.

Remark 7.13. Surprisingly enough, not much else can be said, namely if the quantities that we estimate give consistent estimators. As we have seen before, this may as well not be the case.

7.6 Regression

Simple linear regression model is as follows. We have $X_1, \dots, X_n$ and $(Y_1, \dots, Y_n)$ i.i.d. samples. The hypothesis is that there exist $\varepsilon_1, \dots, \varepsilon_n$ independent with $E\varepsilon_i = 0$ and $\mathrm{Cov}(X_i, \varepsilon_i)$ for each i, such that there exist $\beta_0, \beta_1 \in \mathbb{R}$

$$Y_i = \beta_0 + \beta_1 X_i + \varepsilon_i$$

Suppose we have estimators $\widehat{\beta}_1$ and $\widehat{\beta}_0$ typically given by

$$\beta_1 := \widehat{\mathrm{Corr}}(X, Y) = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2}, \qquad \beta_0 := \bar{Y} - \widehat{\beta}_1 \bar{X}$$

The linear model becomes

$$\widehat{Y}_i := \widehat{\beta}_0 + \widehat{\beta}_1 X_i$$

The residuals are

$$e_i := Y_i - \widehat{Y}_i$$

We want to now use bootstrapping to find variances of β_1, β_0. Simple. We create bootstrap samples $X^*_1, \dots, X^*_n$ and $Y^*_1, \dots, Y^*_n$, but we draw with replacement from $(X_1, Y_1), \dots, (X_n, Y_n)$, so that for each $i, j = 1, \dots, n$

$$P(X^*_i = X_j, Y^*_i = Y_j) = \frac{1}{n}$$

If we repeat this procedure B times we get bootstrap samples (X^{*j}_i) for every $i = 1, \dots, n$, $j = 1, \dots, B$. For each sample, find $\widehat{\beta}^{*j}_0$ and $\widehat{\beta}^{*j}_1$, and now we have empirical distributions. We can compute $\mathrm{Var}(\beta^*_i)$ as well as MSE, and using the quantiles of $\widehat{\beta}^*_i$ we can even construct a confidence interval:

$$[2\widehat{\beta}_i - q^*_{i,1-\alpha/2}, 2\widehat{\beta}_i - q^*_{i,\alpha/2}]$$

8 Bayesian statistics

Definition 8.1. Let (S, Σ, P) be a probability space. Let $A, B \in \Sigma$ be events. Then the **conditional probability** of A given B (when $P(B) > 0$) is given by

$$P(A|B) := \frac{P(A \cap B)}{P(B)}$$

Theorem 8.1 (Bayes' theorem)**.** *Let (S, Σ, P) be a probability space. Let $A, B \in \Sigma$ be events, such that $P(A) > 0$ and $P(B > 0)$, then*

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}$$

Definition 8.2. Let X, Y be random variables with densities f_X, f_Y respectively, and joint density $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$. We define the **conditional density** of X given Y as a function $f_{X|Y} : D \to \mathbb{R}$ such that

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

where $D \subseteq \mathbb{R}^2$ is the domain where $(x, y) \mapsto f_Y(y)$ is nonzero.

Theorem 8.2 (Bayes' theorem for densities)**.** *Let X, Y be random variables with densities f_X, f_Y respectively, and joint density $f_{X,Y} : \mathbb{R}^2 \to \mathbb{R}$. Then for every $x, y \in \mathbb{R}$ such that $f_X(x) \neq 0$ and $f_Y(y) \neq 0$ we have*

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{f_X(x)}$$

Remark 8.1. Notice that very importantly,

$$f_X(x) = \int_{\mathbb{R}} f_{X|Y}(x|y)f_Y(y)\,dy$$

such that we can write Bayes' theorem as

$$f_{Y|X}(y|x) = \frac{f_{X|Y}(x|y)f_Y(y)}{\int_{\mathbb{R}} f_{X|Y}(x|y)f_Y(y)\,dy}$$

which immediately makes it clear why it is a valid probability density.

Let $X_1, \ldots, X_n \sim P_\theta$ be an i.i.d. sample, for which θ is unknown but $\Theta \subseteq \mathbb{R}$, the parameter space, is known, and must be a compact set. Suppose that θ was drawn from a distribution with density π, called the **prior**. Then the **posterior** given data is the conditional density $\pi(\cdot|X_1, \ldots, X_n)$. It can be computed through:

$$\pi(\theta|X_1, \ldots, X_n) = \frac{f_{X|\theta}(X_1, \ldots, X_n|\theta)\pi(\theta)}{\int_\Theta f_{X|\theta}(X_1, \ldots, X_n|\theta)\pi(\theta)\,d\theta}$$

The posterior gives you in a sense the likelihood of θ given our measurements.

Example 8.1. If the prior is chosen uniformly, let's say $\pi(\theta) = \mathbf{1}\{0 \leq \theta \leq 1\}$, we get

$$\pi(\theta|X_1, \ldots, X_n) = \frac{f_{X|\theta}(X_1, \ldots, X_n|\theta)\pi(\theta)}{\int_\Theta f_{X|\theta}(X_1, \ldots, X_n|\theta)\pi(\theta)\,d\theta} = \frac{f_{X|\theta}(X_1, \ldots, X_n|\theta)}{\int_0^1 f_{X|\theta}(X_1, \ldots, X_n|\theta)\,d\theta} = f_{X|\theta}(X_1, \ldots, X_n|\theta)$$

i.e. we get our familiar friend the likelihood back for the posterior.

Definition 8.3. Let $\alpha, \beta > 0$. We say that some random variable X with density f is **beta distributed** if

$$f(x) = \frac{x^{\alpha-1}(1-x)^{\beta-1}}{\int_0^1 y^{\alpha-1}(1-y)^{\beta-1}\,dy}\mathbf{1}\{0 \leq x \leq 1\}$$

We write $X \sim \text{Beta}(\alpha, \beta)$.

Example 8.2. Let $Y \sim U[0,1]$, let $X_1, \dots, X_n \sim U[0,1]$ be i.i.d. and let

$$Z := \sum_{i=1}^{n} \mathbf{1}\{X_i \le Y\}$$

Then what is the distribution of $Z|Y$? If we know, then we can compute an updated distribution $Y|Z$ for Y, that gives you more information about Y given Z. Since X_i are chosen uniformly, we know that $X_i|Y = y \sim \mathrm{Ber}(y)$, so the sum is $B(n,y)$. Then

$$\begin{aligned} f_{Y|Z}(y|z) &= \frac{f_{Z|Y}(z|y) f_Y(y)}{\int_{\mathbb{R}} f_{Z|Y}(z|y) f_Y(y)\, dy} \\ &= \frac{\binom{n}{z} y^z (1-y)^{n-z} \mathbf{1}\{0 \le y \le 1\}}{\int_{\mathbb{R}} \binom{n}{z} y^z (1-y)^{n-z} \mathbf{1}\{0 \le y \le 1\} dy} \\ &= \frac{y^z (1-y)^{n-z} \mathbf{1}\{0 \le y \le 1\}}{\int_{\mathbb{R}} y^z (1-y)^{n-z} \mathbf{1}\{0 \le y \le 1\} dy} \end{aligned}$$

which implies that $Y|Z \sim \mathrm{Beta}(Z+1, n+1-Z)$.

Example 8.3. Let's say we have a linear regression model $Y = X\beta + \varepsilon$, where

$$\varepsilon \sim N(0, \sigma^2 \,\mathrm{id})$$

We observe Y, know X and $\sigma^2 > 0$. Can we use Bayesian statistics to find an approximation of β? Suppose that $\beta \sim N(0, \tau^2 \,\mathrm{id})$ with $\tau^2 > 0$, this is the prior. Recall that $Y \sim N(X\beta, \sigma^2 \,\mathrm{id})$. What is the distribution of $\beta|Y$? Well

$$\begin{aligned} f_{Y|\beta}(y,\beta) f_\beta(\beta) &= \frac{1}{\sqrt{2\pi\sigma^n}} \exp\left(-\frac{\|y - X\beta\|^2}{2\sigma^2} \right) f_\beta(\beta) \\ &= \frac{1}{\sqrt{4\pi^2 \sigma^n \tau^n}} \exp\left(-\frac{\|y - X\beta\|^2}{2\sigma^2} - \frac{\|\beta\|^2}{2\tau^2} \right) \\ &\propto \exp\left(-\frac{\tau^2 \|y\|^2 - 2\tau^2 \langle X^T y, \beta \rangle + (\tau^2 + \sigma^2)\|\beta\|^2}{2(\sigma\tau)^2} \right) \\ &\propto \exp\left(-\frac{\tau^2 \|y\|^2 - \langle 2\tau^2 X^T y + (\tau^2 + \sigma^2)\beta, \beta \rangle}{2(\sigma\tau)^2} \right) \\ &\propto \exp\left(-\frac{\|y\|^2 - \langle 2X^T y + (1 + (\sigma/\tau)^2)\beta, \beta \rangle}{2\sigma^2} \right) \end{aligned}$$

8.1 Estimation

Let's say we have a posterior $\pi(\theta|X)$, how do we estimate?

Definition 8.4 (Posterior mean)**.** The **posterior mean estimator** assigns to each measurement X the estimation

$$\widehat{\theta}(X) := \int_\Theta \theta \pi(\theta|X)\, d\theta$$

i.e. the mean of the posterior. woah.

Definition 8.5 (Posterior mode)**.** The MAP or posterior mode estimator assigns to each measurement X the estimation

$$\widehat{\theta}(X) := \operatorname*{argsup}_{\theta \in \Theta} \pi(\theta|X)$$

Definition 8.6. A **loss function** is a metric on Θ without triangle inequality. That is, it is a map $L : \Theta^2 \to [0, \infty)$ such that for all $\theta_1, \theta_2 \in \Theta$ we have

- $L(\theta_1, \theta_2) = 0 \iff \theta_1 = \theta_2$
- $L(\theta_1, \theta_2) = L(\theta_2, \theta_1)$

Definition 8.7. The following loss functions are common:

- Absolute loss: $L(\theta, \widehat{\theta}) := |\theta - \widehat{\theta}|$
- Quadratic loss: $L(\theta, \widehat{\theta}) := \left(\theta - \widehat{\theta}\right)^2$
- All-or-nothing loss: $L(\theta, \widehat{\theta}) := \mathbf{1}\left\{\theta \neq \widehat{\theta}\right\}$

Definition 8.8. The **Bayes risk** of a loss function L given a parameter θ a random variable in Θ and an estimate $\widehat{\theta} \in \Theta$, is defined to be the expectation of the loss L, that is, the Bayes risk $R(\theta, \widehat{\theta})$ is defined by

$$R(\theta, \widehat{\theta}) := E[L(\theta, \widehat{\theta})]$$

Lemma 8.1. *With the quadratic loss function, the estimator that minimizes the Bayes risk is the posterior mean.*

Proof. Notice that

$$R(\theta, \widehat{\theta}) = E[(\widehat{\theta} - \theta)^2] = \int_\Theta (\widehat{\theta} - \theta)^2 \pi(\theta|X)\, d\theta$$

Minimizing amounts to finding a critical point first, so let's do that

$$\frac{d}{d\widehat{\theta}} \int_\Theta (\widehat{\theta} - \theta)^2 \pi(\theta|X)\, d\theta = 2\int_\Theta (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta = 0$$

which gives

$$\int_\Theta \widehat{\theta}\pi(\theta|X)\, d\theta = \widehat{\theta} = \int_\Theta \theta\pi(\theta|X)\, d\theta$$

is the only critical point. The second derivative is positive, so we have a convex function R for fixed θ, which attains a global minimum at the critical point. Notice that said critical point is the posterior mean. □

Lemma 8.2. *For the absolute loss, the best estimator is the posterior median, i.e. the median of $\theta|X$.*

Proof. Suppose that $\Theta = [a, b]$ for $a < b$ (can be generalized for finite unions of compact intervals and even improper integrals by taking limits). We require that $\widehat{\theta} \in (a, b)$, otherwise we get slightly into trouble (but the R is a.s. not minimized then)

$$\begin{aligned} R(\theta, \widehat{\theta}) &= E[|\widehat{\theta} - \theta|] \\ &= \int_\Theta |\widehat{\theta} - \theta|\pi(\theta|X)\, d\theta \\ &= \int_a^{\widehat{\theta}} (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta - \int_{\widehat{\theta}}^b (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta \end{aligned}$$

We differentiate using the Leibnitz rule

$$\begin{aligned} \frac{\partial}{\partial\widehat{\theta}} R(\theta, \widehat{\theta}) &= \frac{d}{d\widehat{\theta}}\left(\int_a^{\widehat{\theta}} (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta - \int_{\widehat{\theta}}^b (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta\right) \\ &= \frac{d}{d\widehat{\theta}} \int_a^{\widehat{\theta}} (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta + \frac{d}{d\widehat{\theta}} \int_b^{\widehat{\theta}} (\widehat{\theta} - \theta)\pi(\theta|X)\, d\theta \\ &= \int_a^{\widehat{\theta}} \pi(\theta|X)\, d\theta + \int_b^{\widehat{\theta}} \pi(\theta|X)\, d\theta = 0 \end{aligned}$$

This clearly gives you the median of $\theta|X$, now what is remaining is to show that it is indeed a minimum, but once again the original function is convex, using the Leibnitz rule again we get a second derivative of $2\pi(\theta|X) > 0$. □

Lemma 8.3. *For discrete posteriors and the all-or-nothing loss, the posterior mode is the Bayes estimator.*

Definition 8.9. Let $\pi(\cdot|X)$ be the posterior. Let Θ be the parameter space. A $(1-\alpha)$ **credible set** C satisfies

$$\int_C \pi(\theta|X)\,d\theta \geq 1-\alpha$$

Notice that this definition depends on data sample X.

Remark 8.2. Compare credible sets with confidence intervals: if C is a credible set then

$$P(\underbrace{\theta}_{\text{an r.v.}} \in C|\ \underbrace{X}_{\text{data dependency}}) \geq 1-\alpha$$

If C is a confidence interval then

$$\inf_{\theta\in\Theta} P_\theta(\underbrace{\theta}_{\text{fixed value}} \in C) \geq 1-\alpha$$

Definition 8.10. Let C be a credible set, we say that the **frequentist coverage** of C given **true value** $\theta \in \Theta$ is $P_\theta(\theta \in C)$. We say that the (overall) frequentist coverage of C is the infimum over all $\theta \in \Theta$.

Example 8.4. Let $\theta \sim N(0,\tau^2)$ be the prior, let $X|\theta \sim N(\theta,1)$, then we find the posterior:

$$\theta|X \sim N\left(\frac{\tau^2}{1+\tau^2}X, \frac{\tau^2}{1+\tau^2}\right)$$

Note that in this case the posterior mean is then $\frac{\tau^2}{1+\tau^2}X$. Note that as $\tau \to \infty$ the posterior mean approaches X, which is an unbiased estimator of θ. Anyway, we find a credible set. We want to find C such that

$$P(\theta \in C|X) = 1-\alpha$$

Suppose C is of the form $\left[\frac{\tau^2}{1+\tau^2}X - c, \frac{\tau^2}{1+\tau^2}X + c\right]$, where c is to be determined, then for standard normal Z

$$\begin{aligned}
P(\theta \in C|X) &= P\left(-c\sqrt{\frac{1+\tau^2}{\tau^2}} \leq Z \leq c\sqrt{\frac{1+\tau^2}{\tau^2}}\right) \\
&= P\left(Z \leq c\sqrt{\frac{1+\tau^2}{\tau^2}}\right) - P(Z \leq -c\sqrt{\frac{1+\tau^2}{\tau^2}}) \\
&= 2\Phi(c\sqrt{\frac{1+\tau^2}{\tau^2}}) - 1 = 1-\alpha \\
\implies \Phi\left(c\sqrt{\frac{1+\tau^2}{\tau^2}}\right) &= 1-\alpha/2 \\
\implies c &= z_{1-\alpha/2}\sqrt{\frac{\tau^2}{1+\tau^2}}
\end{aligned}$$

Notice that this is very very similar to the standard confidence interval, just slightly narrower and centered differently.

9 Time series

9.1 Definitions

Definition 9.1. A **time series model** is a sequence $(X_t)_{t\in\mathbb{N}^*}$ of random variables, that are not necess

Definition 9.2. A time series $(X_t)_{t\in\mathbb{N}^*}$ is said to be **white noise** if for all $i \neq j$ we have $\mathrm{Cov}(X_i, X_j) = 0$, and have the same first two moments. We write $(X_t) \sim \mathrm{WN}(\mu, \sigma^2)$.

Definition 9.3. A **random walk** is a time series that is the sum of i.i.d. random variables.

Definition 9.4. The **autocovariance function** of a timeseries $(X_t)_{t\in\mathbb{N}^*}$ is a function $\gamma_X : \mathbb{N} \times \mathbb{N} \to \mathbb{R}$ such that

$$(s,t) \mapsto \mathrm{Cov}(X_s, X_t) = E[(X_s - \mu_X(s))(X_t - \mu_X(t))]$$

where $\mu_X : \mathbb{N} \to \mathbb{R}$ is the **mean function** of X, given by $t \mapsto E[X_t]$.

9.2 Stationarity

Definition 9.5. A time series is called **stationary** if

- μ_X is constant
- The forward autocovariance is stationary, that is, for all τ, $\gamma_X(t+\tau, t)$ is independent of t / constant

Definition 9.6. A time series is called **strictly stationary** if the c.d.f is invariant under timeshift, that is, for all $\tau \in \mathbb{Z}$, $n \in \mathbb{N}^*$ and $t_1, \ldots, t_n$, we have

$$F_{t_1+\tau,\ldots,t_n+\tau} = F_{t_1,\ldots,t_n}$$

Where

$$F_{t_1,\ldots,t_n}(x_1, \ldots, x_n) := P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n)$$

Theorem 9.1. *The following properties hold*

- *If X_t is strictly stationary, then they all come from the same distribution*
- *If X_t is strictly stationary, then $(X_t, X_{t+h}) = (X_1, X_{1+h})$ for all t, h*
- *Strict stationarity implies stationarity if the second moment of all time series variables X_t is finite*
- *The converse is not true*
- *An i.i.d. timeseries is strictly stationary*

Proof. Consider the case for a certain i, j and $n = 1$, then we get

$$P(X_j \leq x) = F_{i+(j-i)}(x) = F_i(x) = P(X_i \leq x) \implies X_i = X_j$$

This covers the second case as well? Anyway, given this we easily get that μ_X is constant, and because of that also the autocovariance. The converse not being true can be seen by an example but I don't bother. The last property we prove. Let $(X_t)_{t\in\mathbb{N}^*}$ be an i.i.d. sequence of random variables. Let $t_1, \ldots, t_n$, $x_1, \ldots, x_n$ and $\tau \in \mathbb{Z}$ be arbitrary. Then

$$P(X_{t_1+\tau} \leq x_1, \ldots, X_{t_n+\tau} \leq x_n) = \prod_{i=1}^{n} P(X_{t_i+\tau} \leq x_i) = \prod_{i=1}^{n} P(X_{t_i} \leq x_i) = P(X_{t_1} \leq x_1, \ldots, X_{t_n} \leq x_n)$$

□

Definition 9.7. A timeseries (X_t) is said to be q-dependent for some $q \in \mathbb{N}$ if for all $s, t \in \mathbb{N}^*$ we have the implication

$$|s - t| > q \implies X_s, X_t \text{ are independent}$$

Example 9.1. If $X_t = g(Z_t, Z_{t-1}, \ldots, Z_{t-q})$ for some i.i.d. noise series (Z_t), this is q-dependent. It is however also strictly stationary by the i.i.d. assumption.

9.3 Gaussian processes

Definition 9.8. A timeseries process (X_t) is said to be **Gaussian** if for every $t_1, \dots, t_n \in \mathbb{N}^*$, we have that $X := (X_{t_1}, \dots, X_{t_n})$ is multivariate normal.

Lemma 9.1. *Let (X_t) be Gaussian, then*

$$X_t \text{ stationary} \implies X_t \text{ strictly stationary}$$

In particular, we already had the reverse implication, so we have equivalence.

Proof. Let X_t be Gaussian and stationary. Let $t_1, \dots, t_n \in \mathbb{N}^*$, $\tau \in \mathbb{Z}$, and $x_1, \dots, x_n \in \mathbb{R}$ be arbitary. We want to show that

$$F_{t_1,\dots,t_n} = F_{t_1+\tau,\dots,t_n+\tau}$$

Consider that $X := (X_{t_1}, \dots, X_{t_n}) \sim N(\mu, \Sigma)$ for certain $\mu \in \mathbb{R}^n$ and $\Sigma \in \mathbb{R}^{n\times n}$ a covariance matrix. If we can show that $\widetilde{X} := (X_{t_1+\tau}, \dots, X_{t_n+\tau}) \sim N(\mu, \Sigma)$, which is already by the Gaussian assumption multivariate normal (so it suffices to show that mean and covariance coincide), we are done. Consider that

$$\mu = (EX_{t_1}, \dots, EX_{t_n}) = (\mu_X(t_1), \dots, \mu_X(t_n)) = (\widetilde{\mu}, \dots, \widetilde{\mu})$$

for a certain $\widetilde{\mu} \in \mathbb{R}$ by weak stationarity. The same reasoning can be used to conclude that $EX = E\widetilde{X}$. Now is just suffices to show that the covariance matrix coindices. Consider that

$$\Sigma^i_j = \mathrm{Cov}(X_{t_i}, X_{t_j}) = \gamma_X(t_i, t_j) = \gamma_X(t_j + (t_i - t_j), t_j)$$

Let $\widetilde{\Sigma} := \mathrm{Cov}(\widetilde{X})$, then

$$\widetilde{\Sigma}^i_j = \mathrm{Cov}(X_{t_i+\tau}, X_{t_j+\tau}) = \gamma_X(t_i + \tau, t_j + \tau) = \gamma_X(t_j + \tau + (t_i - t_j), t_j + \tau) = \gamma_X(t_j + (t_i - t_j), t_j) = \Sigma^i_j$$

so $\widetilde{\Sigma} = \Sigma$ and we are done. □

9.4 Auto-things and linear processes

Definition 9.9. Let (X_t) be stationary, then we define the function $\gamma_X : \mathbb{N} \to \mathbb{R}$ by

$$\tau \mapsto \gamma_X(t + \tau, t) = \mathrm{Cov}(X_{t+\tau}, X_t)$$

which is defined independent of t by stationarity, we might as well take $t = |\tau| + 1$ for well-definedness or smth. We additionally define the **autocorrelation function**, $\rho_X : \mathbb{N} \to [-1, 1]$, by

$$\tau \mapsto \frac{\gamma_X(\tau)}{\gamma_X(0)}$$

Theorem 9.2. *The following hold*

- $\gamma_X(0) \geq 0$
- *For all $\tau \in \mathbb{Z}$, $|\gamma_X(\tau)| \leq \gamma_X(0)$*
- *For all $\tau \in \mathbb{Z}$, $\gamma_X(\tau) = \gamma_X(-\tau)$*
- *γ_X is positive definite, that is for all $a \in \mathbb{R}^n$ and arbitary n we have*

$$\sum_{i=1}^{n}\sum_{j=1}^{n} a^i \gamma_X(i - j) a^j \geq 0$$

Proof.

- $\gamma_X(0) = \gamma_X(t, t) = \mathrm{Cov}(X_t, X_t) = \mathrm{Var}(X_t) \geq 0$

- Let $\tau \in \mathbb{Z}$, then

$$\begin{aligned}
|\gamma_X(\tau)| &= |\gamma_X(t+\tau, t)| \\
&= |\operatorname{Cov}(X_{t+\tau}, X_t)| \\
&\leq \sqrt{\operatorname{Var}(X_{t+\tau})\operatorname{Var}(X_t)} \\
&= \sqrt{\operatorname{Var}(X_t)\operatorname{Var}(X_t)} \\
&= \operatorname{Var}(X_t) = \gamma_X(0)
\end{aligned}$$

- Let $\tau \in \mathbb{Z}$, then

$$\gamma_X(-\tau) = \gamma_X(t-\tau, t) = \gamma_X(t-\tau+\tau, t+\tau) = \gamma_X(t, t+\tau) = \gamma_X(t+\tau, t) = \gamma_X(\tau)$$

□

Definition 9.10. A time series $(X_t)_{t\in\mathbb{N}}$ is called **a linear process** if there exists white noise $(W_t)_{t\in\mathbb{Z}} \sim \mathrm{WN}(0, \sigma^2)$ and absolutely summable coefficients $(\psi_j)_{j\in\mathbb{Z}}$ and $\mu \in \mathbb{R}$ such that

$$X_t = \mu + \sum_{j\in\mathbb{Z}} \psi_j W_{t-j}$$

Lemma 9.2. *A linear process is weakly stationary with*

$$\mu_X(t) = \mu$$

$$\gamma_X(\tau) = \sigma^2 \sum_{j\in\mathbb{Z}} \psi_j \psi_{\tau+j}$$

Proof. We simply show the formulae, which directly implies weak stationarity. Consider that

$$\begin{aligned}
\mu_X(t) &= E[X_t] \\
&= \mu + \sum_{j\in\mathbb{Z}} \psi_j E[W_{t-j}] = \mu
\end{aligned}$$

Let $t \in \mathbb{N}$ and $\tau \in Z$, then

$$\begin{aligned}
\gamma_X(\tau) &= \gamma_X(t+\tau, t) \\
&= \operatorname{Cov}(X_{t+\tau}, X_t) \\
&= \operatorname{Cov}\left(\mu + \sum_{j\in\mathbb{Z}} \psi_j W_{t+\tau-j}, \mu + \sum_{j\in\mathbb{Z}} \psi_j W_{t-j}\right) \\
&= \operatorname{Cov}\left(\sum_{i\in\mathbb{Z}} \psi_i W_{t+\tau-i}, \sum_{j\in\mathbb{Z}} \psi_j W_{t-j}\right) \\
&= \sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \psi_i \psi_j \operatorname{Cov}(W_{t+\tau-i}, W_{t-j}) \\
&= \sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \psi_{\tau+i} \psi_j \operatorname{Cov}(W_{t-i}, W_{t-j}) \\
&= \sum_{i\in\mathbb{Z}} \sum_{j\in\mathbb{Z}} \psi_{\tau+i} \psi_j \operatorname{Var}(W_{t-i}) \delta_j^i \\
&= \sigma^2 \sum_{i\in\mathbb{Z}} \psi_{\tau+i} \psi_i
\end{aligned}$$

□

9.5 Estimation on stationary processes

The sample mean of a sample $X_1, \dots, X_n$ of a stationary timeseries

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

is unbiased for $\mu := \mu_X(t)$, this is always true if a sample has the same first moment. We can compute the variance:

$$\begin{aligned}
\operatorname{Var}(\bar{X}) &= \operatorname{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\operatorname{Var}\left(\sum_{i=1}^{n} X_i\right) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\operatorname{Cov}(X_i, X_j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i,j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(j+(i-j),j) \\
&= \frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1}^{n}\gamma_X(i-j) \\
&= \frac{1}{n^2}\left(n\gamma_X(0) + 2\sum_{i=1}^{n-1}(n-i)\gamma_X(i)\right)
\end{aligned}$$

Now if the timeseries is uncorrelated, we get that $\operatorname{Var}(\bar{X}) = \frac{\gamma_X(0)}{n}$ which is neat, very similar to the normal estimator. We can show that

$$\lim_{n\to\infty} n\operatorname{Var}(\bar{X}_n) = \sum_{j\in\mathbb{Z}}\gamma_X(j)$$

so that we have some sort of asymptotic behaviour, and indeed the normal variance as the sample size increases goes to 0. Now if we dealt with a **linear process** we can look at

$$\sum_{\tau\in\mathbb{Z}}\gamma_X(\tau) = \sum_{\tau\in\mathbb{Z}}\sum_{j\in\mathbb{Z}}\sigma^2\psi_j\psi_{\tau+j} = \sigma^2\left(\sum_{j\in\mathbb{Z}}\psi_j\right)^2$$

which gives rise to the theorem

Theorem 9.3. *Let (X_t) be a linear process whose coefficients sum to a nonzero real number, then distributionally*

$$\sqrt{n}(\bar{X}_n - \mu) \to N(0, V)$$

where

$$V := \sigma^2\left(\sum_{j\in\mathbb{Z}}\psi_j\right)^2$$