

Introduction to Statistics and Data Analysis

Niels O. Nygaard

1 Statistical Analysis

We shall consider a dataset consisting of monthly prices of the stock of Intel Corp. from Jan. 1 1990 to Jan 31, 2008. The symbol of this stock is INTC and the data can be downloaded from Yahoo. We import this data into MATLAB. Next we turn the price data into return data by taking the *log* of the prices and taking successive differences. This gives us 216 months=18 years of returns, we group the data into 18 columns of 12 monthly returns. At this point you should watch the accompanying video to see how to do this.

Exercise 1.1 *Watch Video 1 and do the same for the stock of Microsoft (symbol MSFT)*

We are asking the following question: is the average monthly return the same from year to year?

1.1 Distributions and Densities

In order to analyze these data and try to answer this question we need to introduce some concepts from Probability Theory and Statistics.

The idea is that the returns on a stock is a function of many different factors and if we knew what the values of these factors will be next month we could predict the return on the stock. We don't really know what these factors are but let us just define Ω to be the space of all the possible values of these factors and assume we have a probability measure P on this space i.e. there are a certain collection of subsets of Ω , called events, and to each event A we can associate a probability $P(A)$ which is a number $0 \leq P(A) \leq 1$ satisfying a number of reasonable axioms such as $P(\Omega) = 1$ and

the probability of a countable disjoint union of events is the sum of the probabilities of the events etc.

Let $R : \Omega \rightarrow \mathbb{R}$ denote the monthly return on the stock as a function of the values of the factors and let us for each real number $t \in \mathbb{R}$ consider the set $\{\omega \in \Omega | R(\omega) < t\} = R^{-1}(] - \infty, t[)$. We shall assume that this subset is an event. A function $X : \Omega \rightarrow \mathbb{R}$ such that $X^{-1}(] - \infty, x[)$ is an event for every $t \in \mathbb{R}$ is called a random variable. Thus we assume that R is a random variable.

Definition 1.1.1 *Let X be a random variable. Consider the function $\Psi_X : \mathbb{R} \rightarrow [0, 1]$ defined by $\Psi_X(t) = P(X^{-1}(] - \infty, t[))$. This function is called the cumulative distribution function (cdf) of the random variable X*

Thus for each $x \in \mathbb{R}$, $\Psi_X(x)$ is the probability that the value of X is $< x$. We often abbreviate this event as $\{X < x\}$.

Definition 1.1.2 *The probability density function (pdf) is the derivative of the cdf $\phi_X = \Psi'_X$.*

Remark 1.1 *Remark that $\Psi_X(x) = \int_{-\infty}^x \phi_X(t)dt$ and that $P(a \leq X < b) = \int_a^b \phi_X(t)dt$ so $\int_{-\infty}^{\infty} \phi_X(t)dt = 1$*

Exercise 1.2 *Watch Video 2 and use the disttool to view the cdfs and pdfs of all the distributions on the drop-down menu and get a field for how changing parameters change the shape of the graphs*

The most famous distribution is the *Normal* or *Gaussian*. The pdf for the Normal distribution with *mean* μ and *variance* σ^2 is given by

$$\phi(t) = \frac{1}{\sqrt{2\sigma^2\pi}} \exp\left(-\frac{(t - \mu)^2}{2\sigma^2}\right)$$

The cdf is then given by

$$\Psi(x) = \int_{-\infty}^x \phi(t)dt = \frac{1}{\sqrt{2\sigma^2\pi}} \int_{-\infty}^x \exp\left(-\frac{(t - \mu)^2}{2\sigma^2}\right)dt$$

there is no closed form expression for the cdf

If X is a random variable whose distribution is normal with mean μ and variance σ^2 we use the notation $X \sim N(\mu, \sigma)$.

Definition 1.1.3 Let X be a random variable with pdf ϕ_X . The expectation or mean of X is

$$\mathbb{E}(X) = \int_{-\infty}^{\infty} t\phi_X(t)dt$$

The variance of X is

$$\text{var}(X) = \mathbb{E}(X - \mathbb{E}(X))^2$$

Lemma 1.1.1 Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a strictly increasing differentiable function. Let $X : \Omega \rightarrow \mathbb{R}$ be a random variable with pdf ϕ_X . Let $f(X)$ denote the composite function $f \circ X : \Omega \rightarrow \mathbb{R}$. Then $\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(t)\phi_X(t)dt$

Proof: Assume f is an increasing function then $f(X) < f(t) \Leftrightarrow X < t$ hence $\Psi_{f(X)}(f(t)) = \Psi_X(t)$. Taking derivatives we get $\phi_X(t) = \Psi'_X(t) = \Psi'_{f(X)}(f(t))' = \Psi'_{f(X)}(f(t))f'(t) = \phi_{f(X)}(f(t))f'(t)$. Now $\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} u\phi_{f(X)}(u)du$, changing variables in the integral $u = f(t)$ we get $du = f'(t)dt$ and so $\mathbb{E}(f(X)) = \int_{-\infty}^{\infty} f(t)\phi_{f(X)}(f(t))f'(t)dt = \int_{-\infty}^{\infty} f(t)\phi_X(t)dt$

Remark 1.2 This formula holds under much more general conditions, it is not necessary for f to be differentiable or increasing.

Corollary 1.1.1 Let X be a random variable and let $\mu = \mathbb{E}(X)$. Then $\text{var}(X) = \int_{-\infty}^{\infty} (t - \mu)^2\phi_X(t)dt$

Proof: By definition $\text{var}(X) = \mathbb{E}((X - \mu)^2)$. Now apply the preceding lemma (and the remark following it) with $f(t) = (t - \mu)^2$

Exercise 1.3 Show that if $X \sim N(\mu, \sigma^2)$ then $\mathbb{E}(X) = \mu$ and $\text{var}(X) = \sigma^2$

Lemma 1.1.2 Let X be a random variable with $X \sim N(\mu, \sigma^2)$. Then the random variable $\frac{X - \mu}{\sigma} \sim N(0, 1)$

Remark 1.3 The distribution $N(0, 1)$ is known as the standard normal distribution

Proof: We have $P(\frac{X - \mu}{\sigma} < x) = P(X < \mu + \sigma x) = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu + \sigma x} \exp(-\frac{(t - \mu)^2}{2\sigma^2})dt$.

Now change variables $u = \frac{t - \mu}{\sigma}$. Then $du = \frac{dt}{\sigma}$ and when t runs between $-\infty$ and $\mu + \sigma x$, u runs between $-\infty$ and x . The integral then becomes

$\frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\mu+\sigma x} \exp(-\frac{(t-\mu)^2}{2\sigma^2})dt = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^x \exp(-\frac{u^2}{2})du$. This is the standard normal distribution.

Assume now that we have a random variable X and we have observed a value of X say x . We hypothesise that X has a distribution Ψ_X . Does this observed value x support our hypothesis or does it go against it. The idea of statistical hypothesis testing is that if the value x is extremely unlikely i.e. if $P(X \geq x)$ or $P(X \leq x)$ are very small i.e. $\int_x^\infty \phi_X(t)dt$ and $\int_{-\infty}^x \phi_X(t)dt$ are very small then the "sample" value x would tend to contradict our hypothesis.

Example 1.1 *We have a random variable X and our hypothesis is $X \sim N(0, 1)$. We observe a value of $x = -2$. Does this affirm or dis-affirm our hypothesis. Under the hypothesis the probability that $X \leq -2$ is 0.0047 i.e. the probability that a random value of X would be ≤ -2 is exceedingly small so an observed value of -2 would be extremely unlikely under our hypothesis and we would tend to reject our hypothesis.*

If our hypothesis instead is $X \sim N(-1, 0)$ then the probability that $X \leq -2$ is 0.1573 and it is not so unlikely to find a random value ≤ -2 . In this case we might not reject our hypothesis.

Example 1.2 *We hypothesise $X \sim N(0, 1)$ and assume we have observed values $-2, 0.1, -0.19, 1.5, 3, -0.05, .04$. Do these values confirm or contradict the hypothesis?*

Here the idea is that we consider each of these values as a random value of variables X_1, X_2, \dots, X_7 each $\sim N(0, 1)$ and the average $\frac{-2+0.1-0.19+1.5+3-.05+.04}{7} = 0.9143$ as a value of $Y = \frac{X_1 + X_2 + \dots + X_7}{7}$. As we shall see shortly $Y \sim N(0, 1/\sqrt{7})$ and so the probability $P(Y \geq 0.9143) = 0.000624$ which is exceedingly small, thus these data would tend to contradict our hypothesis.

What has happened in this case is that under the hypothesis, the average becomes a value of a random variable with smaller variance and hence the average must be closer to the mean to confirm the hypothesis. Thus the more values we have observed the closer to the mean the average of these values must be in order to not reject the hypothesis.

Remark 1.4 *Usually we will set a level α at which we reject the hypothesis i.e. if $P(X \leq x) < 1 - \alpha$ we reject. The number α is called the confidence*

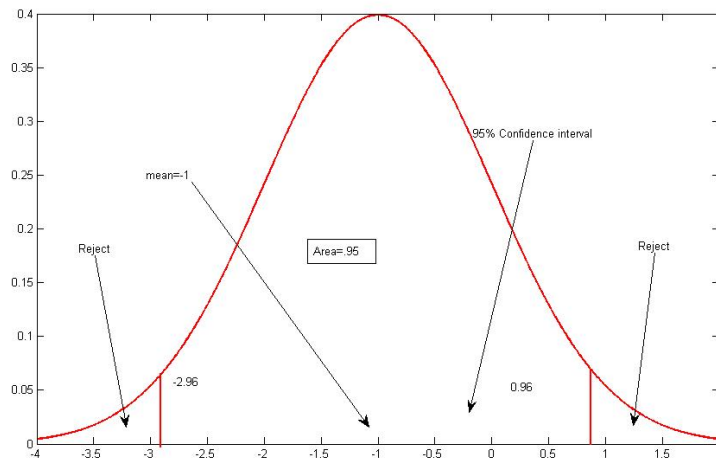


Figure 1: Confidence interval

level and is typically 95% or 99%. We can mark an interval $I = [\mu - a, \mu + a]$ such that $\int_{\mu-a}^{\mu+a} \phi_X(t)dt = \alpha$ called the confidence interval (this is in case the density function is symmetric around μ). If the value x falls outside the confidence interval I we reject at level α . Remark that if we increase the level we make the interval wider hence will be less likely to reject.

Consider now two random variables $X_1, X_2 : \Omega \rightarrow \mathbb{R}$. We define the joint distribution $\Psi_{X_1, X_2} : \mathbb{R}^2 \rightarrow [0, 1]$ by $\Psi_{X_1, X_2}(x_1, x_2) = P(X_1 < x_1, X_2 < x_2)$ i.e. the probability that both $X_1 < x_1$ and $X_2 < x_2$. The joint density is defined by

$$\phi_{X_1, X_2}(t_1, t_2) = \frac{\partial^2}{\partial x_1 \partial x_2} \Psi_{X_1, X_2}(t_1, t_2)$$

so we have

$$\Psi_{X_1, X_2}(x_1, x_2) = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2$$

Definition 1.1.4 We define the covariance between two random variables by $cov(X_1, X_2) = \mathbb{E}(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))$. Thus the covariance is a measure of the tendency of the random variables to move in the same or opposite directions: if the covariance is positive they will tend to "move" in the same direction and if it is negative, in opposite directions. Remark that $cov(X, X) = var(X)$

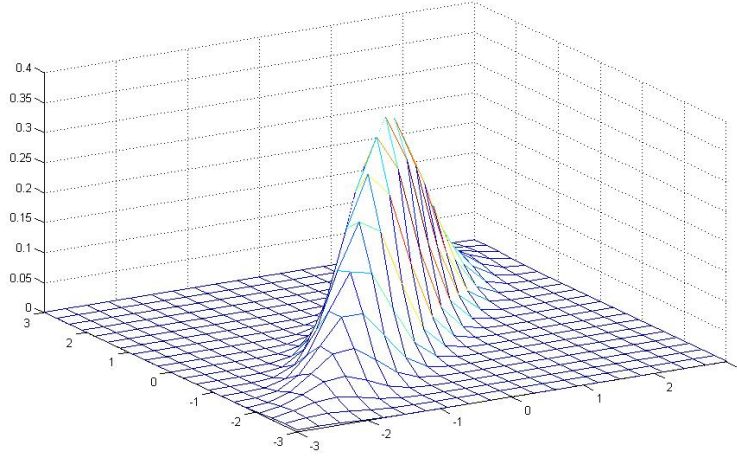


Figure 2: 2-dimensional normal density

Definition 1.1.5 If X_1, X_2, \dots, X_n are random variables the the covariance matrix is the matrix whose i, j 'th entry is $\text{cov}(X_i, X_j)$. Remark that the covariance matrix is symmetric since obviously $\text{cov}(X_i, X_j) = \text{cov}(X_j, X_i)$ and the diagonal terms are the variances.

Example 1.3 Let C be an $n \times n$ symmetric, positive definite matrix. Let $\underline{\mu}$ be a vector in \mathbb{R}^n . The n -variable normal density with mean $\underline{\mu}$ and covariance matrix is the function

$$\phi(x_1, x_2, \dots, x_n) = \phi(\underline{x}) = \frac{1}{(2\pi)^{N/2}(\det C)^{1/2}} \exp\left(-\frac{1}{2}(\underline{x} - \underline{\mu}) \cdot C^{-1} \cdot (\underline{x} - \underline{\mu})\right)$$

Lemma 1.1.3 Let $f(x_1, x_2, \dots, x_n)$ be a function of n variables, and let X_1, X_2, \dots, X_n be random variables with joint density $\phi_{X_1, X_2, \dots, X_n}$. Then the expectation of the random variable $f(X_1, X_2, \dots, X_n)$ is given by

$$\mathbb{E}(f(X_1, X_2, \dots, X_n)) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} f(t_1, \dots, t_n) \phi_{X_1, \dots, X_n}(t_1, \dots, t_n) dt_1 \dots dt_n$$

Example 1.4 As an application of this Lemma consider $\text{cov}(X_1, X_2)$. By definition this is equal to $\mathbb{E}((X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2)))$. Let $\mu_1 = \mathbb{E}(X_1)$ and $\mu_2 = \mathbb{E}(X_2)$ and put $f(x_1, x_2) = (x_1 - \mu_1)(x_2 - \mu_2)$ then $\text{cov}(X_1, X_2) = \mathbb{E}(f(X_1, X_2)) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (t_1 - \mu_1)(t_2 - \mu_2) \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2$

If we have the joint pdf we can recover the individual pdfs as the *marginal* pdfs i.e.

$$\phi_{X_i}(x) = \int_{-\infty}^{\infty} \dots \int_{-\infty}^{\infty} \phi_{X_1, \dots, X_n}(t_1, \dots, t_{i-1}, x, t_{i+1}, \dots, t_n) dt_1 \dots dt_{i-1} dt_{i+1} \dots dt_n$$

Definition 1.1.6 Consider random variables $X_1, X_2, \dots, X_m, \dots, X_n$. The conditional density function is defined by

$$\phi_{X_1, X_2, \dots, X_m | X_{m+1}, \dots, X_n}(t_1, t_2, \dots, t_m | t_{m+1}, \dots, t_n) = \frac{\phi_{X_1, \dots, X_m, \dots, X_n}(t_1, \dots, t_m, \dots, t_n)}{\phi_{X_{m+1}, \dots, X_n}(t_{m+1}, \dots, t_n)}$$

Two random variables X_1, X_2 are said to be independent if $\phi_{X_1 | X_2}(t_1 | t_2) = \phi_{X_1}(t_1)$ or equivalently if $\phi_{X_1, X_2}(t_1, t_2) = \phi_{X_1}(t_1)\phi_{X_2}(t_2)$

Definition 1.1.7 Let A and B be events. The conditional probability $P(A|B)$ is defined by $P(A|B) = \frac{P(A \cap B)}{P(B)}$. Two events are independent if $P(A|B) = P(A)$ or equivalently $P(A \cap B) = P(A)P(B)$

Definition 1.1.8 For random variables X_1 and X_2 we can define the conditional distribution by $\Psi_{X_1 | X_2}(x_1 | x_2) = P(X_1 < x_1 | X_2 < x_2)$

Proposition 1.1.1 Random variables X_1, X_2 are independent if and only if for all x_1, x_2 the events $\{X_1 < x_1\}$ and $\{X_2 < x_2\}$ are independent

Proof: If $\phi_{X_1, X_2}(t_1, t_2) = \phi_{X_1}(t_1)\phi_{X_2}(t_2)$ we have

$$\begin{aligned} \Psi_{X_1, X_2}(x_1, x_2) &= \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2 = \int_{-\infty}^{x_1} \int_{-\infty}^{x_2} \phi_{X_1}(t_1)\phi_{X_2}(t_2) dt_1 dt_2 \\ &= \int_{-\infty}^{x_1} \phi_{X_1}(t_1) dt_1 \int_{-\infty}^{x_2} \phi_{X_2}(t_2) dt_2 = \Psi_{X_1}(x_1)\Psi_{X_2}(x_2) \end{aligned}$$

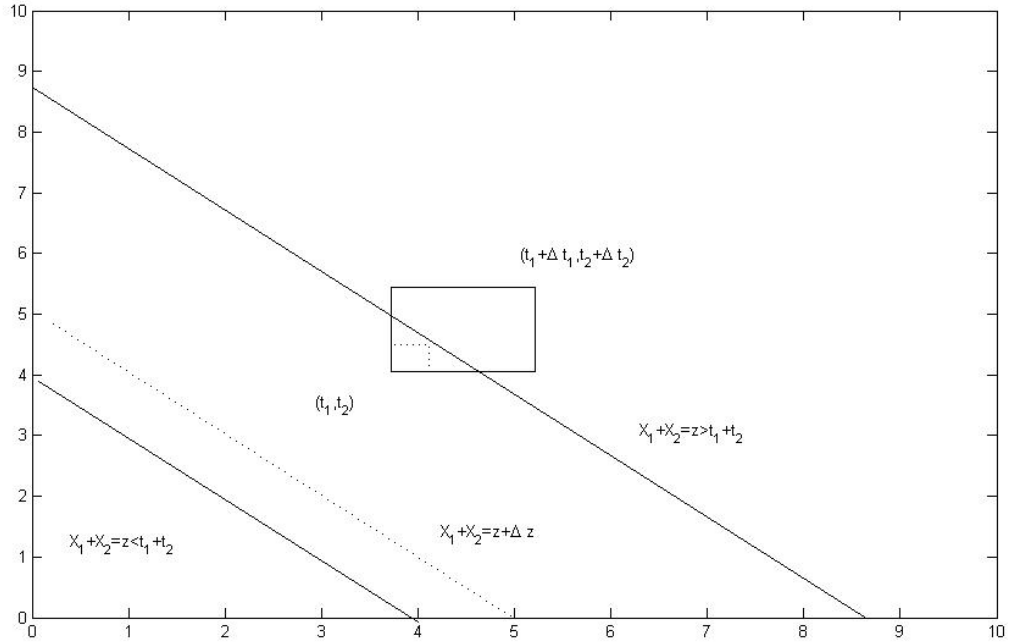
To go the other way we have $\phi_{X_1, X_2} = \frac{\partial^2}{\partial x_1 \partial x_2} \Psi_{X_1, X_2}$. If $\Psi_{X_1, X_2}(x_1, x_2) = \Psi_{X_1}(x_1)\Psi_{X_2}(x_2)$ we have $\frac{\partial}{\partial x_2} \Psi_{X_1, X_2}(x_1, x_2) = \Psi_{X_1}(x_1) \frac{d}{dx_2} \Psi_{X_2}(x_2) = \Psi_{X_1}(x_1)\phi_{X_2}(x_2)$. Now taking the partial derivative with respect to x_1 proves the other direction

Consider random variables X_1, X_2 and form the random variable $Z = X_1 + X_2$. What is the density of Z given the densities of X_1 and X_2 ? Remark

that $\phi_X(t)\Delta t$ when Δt is small can be interpreted as an approximation to the probability that $X \in [t, t+\Delta t]$ since $\phi_X(t) = \lim_{\Delta t \rightarrow 0} \frac{P(X < t + \Delta t) - P(X < t)}{\Delta t} = \lim_{\Delta t \rightarrow 0} \frac{P(t \leq X < t + \Delta t)}{\Delta t}$. Similarly $\phi_{X_1, X_2}(t_1, t_2)\Delta t_1\Delta t_2$ can be interpreted as (a close approximation to) the probability that (X_1, X_2) is in the rectangle $[t_1, t_1 + \Delta t_1] \times [t_2, t_2 + \Delta t_2]$.

Now

$$\begin{aligned} \phi_{Z|X_1, X_2}(z|t_1, t_2)\Delta z &= \frac{\phi_{Z, X_1, X_2}(z, t_1, t_2)\Delta z\Delta t_1\Delta t_2}{\phi_{X_1, X_2}(t_1, t_2)\Delta t_1\Delta t_2} \\ &\sim \frac{P(z \leq Z < z + \Delta z, t_1 \leq X_1 < t_1 + \Delta t_1, t_2 \leq X_2 < t_2 + \Delta t_2)}{P(t_1 \leq X_1 < t_1 + \Delta t_1, t_2 \leq X_2 < t_2 + \Delta t_2)} \end{aligned}$$



The figure shows that if $z \neq t_1 + t_2$ we can take $\Delta z, \Delta t_1, \Delta t_2$ small enough that $\{z \leq Z < z + \Delta z\} \cap \{(X_1, X_2) \in [t_1, t_1 + \Delta t_1] \times [t_2, t_2 + \Delta t_2]\} = \emptyset$ and so the numerator is 0.

If $z = t_1 + t_2$ $\{(X_1, X_2) \in [t_1, t_1 + \Delta t_1] \times [t_2, t_2 + \Delta t_2]\} \subset \{z \leq Z < z + \Delta z\}$ when Δt_1 and Δt_2 are sufficiently small. Hence the fraction = 1. It follows that $\phi_{Z|X_1, X_2}(z, t_1, t_2)\Delta z = \begin{cases} 0 & z \neq t_1 + t_2 \\ 1 & z = t_1 + t_2 \end{cases}$ Letting $\Delta z \rightarrow 0$ we get that $\phi_{Z|X_1, X_2}(z|t_1, t_2) = \delta(z - (t_1 + t_2))$, the Dirac delta function.

Now we can compute the density of Z : write ϕ_Z as the marginal density of ϕ_{Z, X_1, X_2}

$$\begin{aligned} \phi_Z(z) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_{Z, X_1, X_2}(z, t_1, t_2) dt_1 dt_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \phi_{Z|X_1, X_2}(z|t_1, t_2) \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta(z - (t_1 + t_2)) \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2 \end{aligned}$$

Recall that we have the general identity $\int_{-\infty}^{\infty} f(t)\delta(x - t)dt = f(x)$. Applying this identity we get

$$\phi_Z(z) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} \delta((z - t_2) - t_1) \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2 = \int_{-\infty}^{\infty} \phi_{X_1, X_2}(z - t_2, t_2) dt_2$$

Proposition 1.1.2 *If X_1, X_2 are independent $\text{cov}(X_1, X_2) = 0$.*

Proof: We have

$$\begin{aligned} \text{cov}(X_1, X_2) &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (t_1 - \mu_1)(t_2 - \mu_2) \phi_{X_1, X_2}(t_1, t_2) dt_1 dt_2 \\ &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (t_1 - \mu_1)(t_2 - \mu_2) \phi_{X_1}(t_1) \phi_{X_2}(t_2) dt_1 dt_2 \\ &= \int_{-\infty}^{\infty} (t_1 - \mu_1) \phi_{X_1}(t_1) dt_1 \int_{-\infty}^{\infty} (t_2 - \mu_2) \phi_{X_2}(t_2) dt_2 \end{aligned}$$

But

$$\int_{-\infty}^{\infty} (t_1 - \mu_2) \phi_{X_1}(t_1) dt_1 = \int_{-\infty}^{\infty} t_1 \phi_{X_1}(t_1) dt_1 - \mu_1 \int_{-\infty}^{\infty} \phi_{X_1}(t_1) dt_1 = 0$$

because $\int_{-\infty}^{\infty} t_1 \phi_{X_1}(t_1) dt_1 = \mu_1$ and $\int_{-\infty}^{\infty} \phi_{X_1}(t_1) dt_1 = 1$

Remark 1.5 In general the reverse statement is false: $\text{cov}(X_1, X_2) = 0$ does not imply X_1, X_2 independent. It is however true if the joint distribution is normal. Indeed if $\text{cov}(X_1, X_2) = 0$ the covariance matrix is diagonal $C = \begin{pmatrix} \sigma_1^2 & 0 \\ 0 & \sigma_2^2 \end{pmatrix}$ and hence the joint density

$$\phi_{X_1, X_2}(t_1, t_2) = \frac{1}{(2\pi)^{N/2}(\det C)^{1/2}} \exp\left(-\frac{1}{2}(\underline{t} - \underline{\mu}) \cdot C^{-1} \cdot {}^t(\underline{t} - \underline{\mu})\right),$$

where $N = 2$, $\det C = \sigma_1^2 \sigma_2^2$,

$$(\underline{t} - \underline{\mu}) \cdot C^{-1} \cdot {}^t(\underline{t} - \underline{\mu}) = (t_1 - \mu_1, t_2 - \mu_2) \begin{pmatrix} \sigma_1^{-2} & 0 \\ 0 & \sigma_2^{-2} \end{pmatrix} \begin{pmatrix} t_1 - \mu_1 \\ t_2 - \mu_2 \end{pmatrix} = \frac{(t_1 - \mu_1)^2}{\sigma_1^2} + \frac{(t_2 - \mu_2)^2}{\sigma_2^2},$$

is equal to

$$\frac{1}{\sigma_1 \sqrt{2\pi}} \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{(t_1 - \mu_1)^2}{2\sigma_1^2} - \frac{(t_2 - \mu_2)^2}{2\sigma_2^2}\right) = \phi_{X_1}(t_1) \phi_{X_2}(t_2)$$

Definition 1.1.9 Let $\{X_n\}$ be a sequence of random variables and let X be a random variable. We say that $X_n \rightarrow X$ in probability if for all $\varepsilon > 0$, $P(|X_n - X| > \varepsilon) \rightarrow 0$.

We have two very important theorems:

Theorem 1.1.1 (Law of Large Numbers) Let X_1, X_2, \dots be a sequence of independent, identically distributed random variables, each with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Define $Y_n = \frac{X_1 + X_2 + \dots + X_n}{n}$. Then $Y_n \rightarrow \mu$ in probability.

Theorem 1.1.2 The Central Limit Theorem Let X_1, X_2, \dots be a sequence of independent, identically distributed random variables, each with $\mathbb{E}(X_i) = \mu$ and $\text{var}(X_i) = \sigma^2$. Set $Z_n = \frac{(X_1 - \mu) + (X_2 - \mu) + \dots + (X_n - \mu)}{\sqrt{n}}$.

Then $\mathbb{E}(Z_n) = 0$ and $\text{var}(Z_n) = \sigma^2$ and the distribution of Z_n converges to the normal distribution with mean 0 and variance σ^2 for $n \rightarrow \infty$

To illustrate the use of the Law of Large Numbers, consider an experiment where we make some kind of measurement and repeat the identical experiment many times. The measurement is of the form $X_i = \mu + \epsilon_i$ where μ is

the true value of the quantity we are measuring and ϵ_i is an error term that has some distribution with some variance σ^2 and mean 0. If the experiment is done under identical conditions, the distributions of the X_i s are identical. The theorem tells us that the probability that $\frac{X_1 + X_2 + \dots + X_n}{n}$ is different from μ goes to 0. In other words the averages of the measurements converge to μ with probability 1.

Theorem 1.1.3 *Let $X_1 \sim N(\mu_1, \sigma_1^2)$ and $X_2 \sim N(\mu_2, \sigma_2^2)$ be independent random variables. Then $X_1 + X_2 \sim N(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$*

Proof: Since $\mathbb{E}(X_1 + X_2) = \mu_1 + \mu_2$ we have $\mathbb{E}((X_1 - \mu_1) + (X_2 - \mu_2)) = 0$ hence there is no loss of generality by assuming $\mu_1 = \mu_2 = 0$. From the computation above we have

$$\phi_{X_1+X_2}(z) = \int_{-\infty}^{\infty} \phi_{X_1, X_2}(z-t, t) dt$$

. Since X_1, X_2 are independent we have $\phi_{X_1, X_2} = \phi_{X_1} \phi_{X_2}$ and so

$$\begin{aligned} \phi_{X_1+X_2}(z) &= \int_{-\infty}^{\infty} \frac{1}{\sigma_1 \sqrt{2\pi}} \frac{1}{\sigma_2 \sqrt{2\pi}} \exp\left(-\frac{(z-t)^2}{2\sigma_1^2}\right) \exp\left(-\frac{t^2}{2\sigma_2^2}\right) dt \\ &= \frac{1}{\sigma_1 \sigma_2 2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{\sigma_2^2(z-t)^2 + \sigma_1^2 t^2}{2\sigma_1^2 \sigma_2^2}\right) dt \end{aligned}$$

Now we can rewrite

$$\begin{aligned}
\frac{\sigma_2^2(z-t)^2 + \sigma_1^2 t^2}{2\sigma_1^2 \sigma_2^2} &= \frac{\sigma_2^2 z^2 + \sigma_2^2 t^2 - 2\sigma_2^2 zt + \sigma_1^2 t^2}{2\sigma_1^2 \sigma_2^2} \\
&= \frac{(\sigma_1^2 + \sigma_2^2)(t^2 - 2\frac{\sigma_2^2 zt}{\sigma_1^2 + \sigma_2^2}) + \sigma_2^2 z^2}{2\sigma_1^2 \sigma_2^2} \\
&= \frac{(\sigma_1^2 + \sigma_2^2) \left((t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})^2 - \frac{\sigma_2^4 z^2}{(\sigma_1^2 + \sigma_2^2)^2} \right) + \sigma_2^2 z^2}{2\sigma_1^2 \sigma_2^2} \\
&= \frac{(\sigma_1^2 + \sigma_2^2)(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})^2 - \frac{\sigma_2^4 z^2}{(\sigma_1^2 + \sigma_2^2)} + (\sigma_1^2 + \sigma_2^2)\frac{\sigma_2^2 z^2}{(\sigma_1^2 + \sigma_2^2)}}{2\sigma_1^2 \sigma_2^2} \\
&= \frac{(\sigma_1^2 + \sigma_2^2)(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})^2 + \frac{-\sigma_2^4 z^2 + \sigma_1^2 \sigma_2^2 z^2 + \sigma_2^4 z^2}{(\sigma_1^2 + \sigma_2^2)}}{2\sigma_1^2 \sigma_2^2} \\
&= \frac{(\sigma_1^2 + \sigma_2^2)(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})^2}{2\sigma_1^2 \sigma_2^2} + \frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}
\end{aligned}$$

It follows that the integral becomes

$$\begin{aligned}
&\frac{1}{\sigma_1 \sigma_2 2\pi} \int_{-\infty}^{\infty} \exp\left(-\frac{(\sigma_1^2 + \sigma_2^2)(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})^2}{2\sigma_1^2 \sigma_2^2} + \frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right) dt \\
&= \frac{1}{\sigma_1 \sigma_2 2\pi} \exp\left(-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right) \int_{-\infty}^{\infty} \exp\left(-\frac{(\sigma_1^2 + \sigma_2^2)(t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})^2}{2\sigma_1^2 \sigma_2^2}\right) dt
\end{aligned}$$

In the last integral we make the substitution $u = \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{2}\sigma_1 \sigma_2} (t - \frac{\sigma_2^2 z}{\sigma_1^2 + \sigma_2^2})$

then $du = \frac{\sqrt{\sigma_1^2 + \sigma_2^2}}{\sqrt{2}\sigma_1 \sigma_2} dt$ and hence the last integral becomes

$$\frac{\sqrt{2}\sigma_1 \sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \int_{-\infty}^{\infty} \exp(-u^2) du = \frac{\sqrt{2}\sigma_1 \sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \sqrt{\pi}$$

It follows that

$$\phi_{X_1+X_2}(z) = \frac{1}{\sqrt{\sigma_1^2 + \sigma_2^2}\sqrt{2\pi}} \exp\left(-\frac{z^2}{2(\sigma_1^2 + \sigma_2^2)}\right)$$

i.e. $X_1 + X_2 \sim N(0, \sigma_1^2 + \sigma_2^2)$

Remark 1.6 This theorem easily generalizes to a linear combination of independent random variables $X_i \sim N(\mu_i, \sigma_i^2)$, then $a_1X_1 + a_2X_2 + \dots + a_nX_n \sim N(a_1\mu_1 + a_2\mu_2 + \dots + a_n\mu_n, a_1^2\sigma_1^2 + a_2^2\sigma_2^2 + \dots + a_n^2\sigma_n^2)$

Definition 1.1.10 Let $X_1, X_2, \dots, X_n, X_i \sim N(0, 1)$ be independent random variables. Then the distribution of the random variable $Z = X_1^2 + X_2^2 + \dots + X_n^2$ is called the χ^2 distribution with n degrees of freedom (df). The density of the $\chi^2(n)$ distribution is given by the function $\phi_Z(z) = \frac{z^{\frac{n-2}{2}} \exp(-z/2)}{2^{n/2}\Gamma(n/2)}$

Definition 1.1.11 Student's t-distribution with n degrees of freedom is the distribution of a quotient $T = \frac{X}{\sqrt{\frac{Z}{n}}}$ where $X \sim N(0, 1)$ and $Z \sim \chi^2(n)$. The

density function of a $t(n)$ -distribution is given by $\phi_T(t) = \frac{\Gamma(\frac{n}{2})}{\sqrt{n\pi}\Gamma(\frac{n-1}{2})} \left(1 + \frac{t^2}{n}\right)^{-\frac{(n+1)}{2}}$

Definition 1.1.12 The F -distribution with degrees of freedom (n, m) is the distribution of a quotient of random variables $F = \frac{Y/n}{Z/m}$ where $Y \sim \chi^2(n)$ and $Z \sim \chi^2(m)$. The density of the $F(n, m)$ -distribution is given by $\phi_F(x) = \frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \binom{n}{m}^{\frac{n}{2}} \frac{x^{\frac{n-2}{2}}}{(1 + (\frac{n}{m})x)^{\frac{n+m}{2}}}$

Exercise 1.4 Use the disttool to view the cdfs and pdfs of the χ^2 -distribution, the t -distribution and the F -distribution. What happens as the degrees of freedom increase?

1.2 Estimators and Hypothesis Testing

Consider again our imaginary experiment where we perform the experiment and measure some quantity. Performing this many times over we will most

likely get a different result every time even if the experiments are conducted under identical conditions, using identical measuring equipment etc. This is because even the most precise measurements have small random variations. Our model is that the result of the i 'th measurement is of the form $X_i = \mu + \varepsilon_i$ where ε_i is the random error term and μ is the true value of the quantity we are measuring. Suppose now that we hypothesise a theoretical value μ_0 for the quantity we are measuring. Our task is to determine whether our measurements confirm or reject our hypothesis.

Without some assumptions about the distribution of the error terms there is not much we can do statistically.

It is reasonable to assume that the random variables ε_i and ε_j are independent for $i \neq j$ and are identically distributed and also that the mean is 0. If the mean is not 0 we are making a systematic error in our measurement. We will make the assumption that $\varepsilon_i \sim N(0, \sigma^2)$ where σ^2 is known. The assumption of normality can be justified by the Central Limit Theorem but the knowledge of the true variance is not very realistic and we shall see later how to get rid of this assumption.

Under this assumption $X_i \sim N(\mu, \sigma^2)$ and so taking the average we know by the theorem from the last section that $\bar{X} = \frac{X_1 + X_2 + \dots + X_n}{n} \sim N(\mu, \sigma^2/n)$, the Law of Large Numbers tells us that $\bar{X} \rightarrow \mu$ in probability for $n \rightarrow \infty$. The random variable \bar{X} is an *estimator* for the mean.

Assumed we have measured x_1, x_2, \dots, x_n then under our hypothesis the *sample average* $\bar{x} = \frac{1}{n}(x_1 + x_2 + \dots + x_n)$ is a random value of a $N(\mu, \sigma^2/n)$ distributed random variable. If the sample average falls outside a 95% confidence interval we will reject the hypothesis at the 95'th percentage confidence level, otherwise we will not reject it. In practical terms we compute the z -value: $z = \frac{\bar{x} - \mu}{\frac{\sigma}{\sqrt{n}}}$ then $z \sim N(0, 1)$ and we can compute the value of the cdf of the normal distribution with mean 0 and variance 1 (known as the *standard normal distribution*). If the value of the cdf is either very small or very close to 1 we reject. This test is known as the z -test. The hypothesis we are testing is commonly known as the *null-hypothesis*, H_0 . Thus in this case H_0 : X_i is normally distributed with mean μ and variance σ . The *alternate hypothesis*, H_1 is that it is not. Statistical tests are really about testing whether the observed data rejects a given null-hypothesis and not to prove that a given null-hypothesis is true.

Exercise 1.5 Watch Video 3 and after that use MATLAB and the data set Data1 on the web page and test the following hypotheses at the 95'th and the 99'th percentage level

1. The data follows a $N(0.5, 1)$ distribution
2. The data follows a $N(0.7, .1)$ distribution

Next we shall abandon the unrealistic assumption that the variance of the error term is known. This means that we need to find an estimator for the variance. Since the variance is $var(X) = \mathbb{E}((X - \mathbb{E}(X))^2)$ we could use the random variable $S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2$ as an estimator for the variance. Here again the X_i s are independent, identically distributed with mean μ and variance σ^2

Definition 1.2.1 Let Z be an estimator for a quantity α . Then we say that Z is un-biased if the expectation of Z , $\mathbb{E}(Z) = \alpha$ otherwise it is a biased estimator.

Proposition 1.2.1 $\mathbb{E}(\frac{1}{n} \sum (X_i - \bar{X})^2) = \frac{n}{n-1} \sigma^2$, thus S^2 is a biased estimator. To get an unbiased estimator we can take $\frac{1}{n-1} \sum (X_i - \bar{X})^2$

Proof: We have $S^2 = \frac{1}{n} \sum (X_i - \bar{X})^2 = \frac{1}{n} \sum ((X_i - \mu) - (\bar{X} - \mu))^2 = \frac{1}{n} (\sum (X_i - \mu)^2 + n(\bar{X} - \mu)^2 - 2(\bar{X} - \mu) \sum (X_i - \mu))$. Now $\sum (X_i - \mu) = n(\bar{X} - \mu)$ so we get $S^2 = \frac{1}{n} \sum (X_i - \mu)^2 - (\bar{X} - \mu)^2$. Taking expectations we get $\mathbb{E}(S^2) = \frac{1}{n} \sum \mathbb{E}(X_i - \mu)^2 - \mathbb{E}(\bar{X} - \mu)^2 = \frac{1}{n} \sum \sigma^2 - \mathbb{E}(\bar{X} - \mu)^2$. Now $\mu = \mathbb{E}(\bar{X})$ so the last term is just $var(\bar{X})$ and so we get $\mathbb{E}(S^2) = \sigma^2 - var(\bar{X})$. This already shows that $\mathbb{E}(S^2) \neq \sigma^2$ hence S^2 is a biased estimator.

To compute $var(\bar{X})$ write

$$\begin{aligned} \mathbb{E}(\bar{X} - \mu)^2 &= \mathbb{E}((\frac{1}{n} \sum X_i - \mu)^2) = \mathbb{E}((\frac{1}{n} \sum X_i - \mu)^2) \\ &= \mathbb{E} \left(\frac{1}{n^2} \sum_{i,j} (X_i - \mu)(X_j - \mu) \right) = \frac{1}{n^2} \sum_{i,j} \mathbb{E}((X_i - \mu)(X_j - \mu)) \end{aligned}$$

Now $\mathbb{E}((X_i - \mu)(X_j - \mu)) = \text{cov}(X_i, X_j) = \begin{cases} 0 & \text{if } i \neq j \\ \text{var}(X_i) = \sigma^2 & \text{if } i = j \end{cases}$

Hence $\text{var}(\bar{X}) = \frac{1}{n^2} \sum_i \sigma^2 = \frac{\sigma^2}{n}$ and thus $\mathbb{E}(S^2) = \sigma^2 - \frac{1}{n} \sigma^2 = \frac{n-1}{n} \sigma^2$

Let $s^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$. We define *the test statistic* T by $T = \frac{\bar{X} - \mu}{\sqrt{\frac{s^2}{n}}}$.

To use this test statistic we need to know its distribution. We already know that the numerator is standard normal so it is the denominator we have to investigate. Consider first $\sum \left(\frac{X_i - \mu}{\sigma} \right)^2$. This is a sum of n squares of standard normals and hence is $\chi^2(n)$. However we don't have $\frac{X_i - \mu}{\sigma}$ but rather $\frac{X_i - \bar{X}}{\sigma}$ so we can't immediately conclude that the denominator is a square root of a $\chi^2(n)$.

Lemma 1.2.1 Consider the vector in \mathbb{R}^n , $\mathbf{1} = (1, 1, \dots, 1)$. Let P denote the orthogonal projection in the direction of $\mathbf{1}$. P takes a vector $\underline{x} = (x_1, x_2, \dots, x_n)$ to the vector $(x_1, x_2, \dots, x_n) - (\bar{x}, \bar{x}, \dots, \bar{x})$ where $\bar{x} = \frac{1}{n}(x_1 +$

$x_2 + \dots + x_n)$. Let $\underline{Z} = \begin{pmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_n \end{pmatrix}$ be a vector of independent standard normal random variables. Then ${}^t \underline{Z} \cdot P \cdot \underline{Z} \sim \chi^2(n-1)$

Proof: Consider the matrix of $P = \begin{pmatrix} 1 - 1/n & -1/n & \dots & -1/n \\ -1/n & 1 - 1/n & \dots & -1/n \\ \vdots & \vdots & \ddots & \vdots \\ -1/n & -1/n & \dots & 1 - 1/n \end{pmatrix}$

Now consider the subspace $V = \{\mathbf{1}\}^\perp$. This is the subspace of vectors $\underline{x} = (x_1, x_2, \dots, x_n)$ such that $\sum x_i = 0$. Let $\underline{q}_1 = \frac{1}{\sqrt{n}} \mathbf{1}$ and let $\underline{q}_2, \underline{q}_3, \dots, \underline{q}_n$ be an orthonormal basis of V i.e. $\|\underline{q}_i\| = 1$ and $\underline{q}_i \perp \underline{q}_j$ for $i \neq j$. Consider the matrix Q where the columns are the vectors $\underline{q}_1, \underline{q}_2, \dots, \underline{q}_n$, then Q is an

orthogonal matrix i.e. ${}^tQ \cdot Q = I_n$ where I_n is the $n \times n$ identity matrix. Remark that $P\underline{x} = \underline{x}$ for any vector in V and $P\underline{1} = \underline{0}$.

Consider the matrix product ${}^tQ \cdot P \cdot Q$. Applying this matrix to the first standard basis vector $\underline{e}_1 = (1, 0, 0, \dots, 0)$ we get $Q\underline{e}_1 = \underline{q}_1$ and $P\underline{q}_1 = \underline{0}$ because \underline{q}_1 is a multiple of $\underline{1}$. For any of the other standard basis vectors $\underline{e}_i = (0, 0, \dots, 1, 0, \dots, 0)$ we have $Q\underline{e}_i = \underline{q}_i \in V$ so $P\underline{q}_i = \underline{q}_i$. Hence $P \cdot Q\underline{e}_i = Q\underline{e}_i$ for $i = 2, 3, \dots, n$. Thus ${}^tQ \cdot P \cdot Q\underline{e}_i = {}^tQ \cdot Q\underline{e}_i = \underline{e}_i$ because

$${}^tQ \cdot Q = I_n. \text{ This shows that } {}^tQ \cdot P \cdot Q = \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix}. \text{ Consider now}$$

the vector of random variables $\underline{Y} = {}^tQ\underline{Z}$. Each of the Y_i 's are of the form $a_{i1}Z_1 + a_{i2}Z_2 + \dots + a_{in}Z_n$ where $\underline{q}_i = (a_{i1}, a_{i2}, \dots, a_{in})$ is the i 'th column in Q . Thus $Y_i \sim N(0, a_{i1}^2 + a_{i2}^2 + \dots + a_{in}^2)$. But $\|\underline{q}_i\| = 1$ so $a_{i1}^2 + a_{i2}^2 + \dots + a_{in}^2 = 1$. It follows that $Y_i \sim N(0, 1)$. Next consider the matrix of random variables

$$\underline{Y} \cdot {}^t\underline{Y} = \begin{pmatrix} Y_1^2 & Y_1Y_2 & \dots & Y_1Y_n \\ Y_1Y_2 & Y_2^2 & \dots & Y_2Y_n \\ \vdots & \vdots & \ddots & \vdots \\ Y_1Y_n & Y_2Y_n & \dots & Y_n^2 \end{pmatrix}. \text{ Taking expectations we get } \mathbb{E}(\underline{Y} \cdot {}^t\underline{Y}) = \begin{pmatrix} \text{var}(Y_1) & \text{cov}(Y_1, Y_2) & \dots & \text{cov}(Y_1, Y_n) \\ \text{cov}(Y_2, Y_1) & \text{var}(Y_2) & \dots & \text{cov}(Y_2, Y_n) \\ \vdots & \vdots & \ddots & \vdots \\ \text{cov}(Y_1, Y_n) & \text{cov}(Y_2, Y_n) & \dots & \text{var}(Y_n^2) \end{pmatrix}.$$

But $\mathbb{E}(\underline{Y} \cdot {}^t\underline{Y}) = \mathbb{E}({}^tQ \cdot \underline{Z} \cdot {}^t\underline{Z} \cdot Q) = {}^tQ \cdot \mathbb{E}(\underline{Z} \cdot {}^t\underline{Z}) \cdot Q$. Since the Z_i 's are independent $\sim N(0, 1)$ the matrix of covariances is the identity matrix I_n . Thus ${}^tQ \cdot \mathbb{E}(\underline{Z} \cdot {}^t\underline{Z}) \cdot Q = {}^tQ \cdot I_n \cdot Q = I_n$ because Q is orthogonal. Thus the matrix of covariances of the Y_i 's is also I_n and so $\text{cov}(Y_i, Y_j) = 0$ for $i \neq j$. Since the Y_i 's are standard normal and their covariances are 0 they are independent.

It follows that $Y_2^2 + Y_3^2 + \dots + Y_n^2 \sim \chi^2(n-1)$.

$$Y_2^2 + Y_3^2 + \dots + Y_n^2 = {}^t\underline{Y} \begin{pmatrix} 0 & 0 & 0 & \dots & 0 \\ 0 & 1 & 0 & \dots & 0 \\ 0 & 0 & 1 & \dots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \dots & 1 \end{pmatrix} \underline{Y} = {}^t\underline{Y} {}^tQ \cdot P \cdot Q\underline{Y}$$

Thus $\underline{Y} = {}^tQ\underline{Z}$ we get $Y_2^2 + \dots + Y_n^2 = \underline{Z}Q \cdot {}^tQ \cdot P \cdot Q {}^tQ\underline{Z} = {}^t\underline{Z} \cdot P \cdot \underline{Z} \sim \chi^2(n-1)$

We shall apply this lemma in the case where $Z_i = \frac{X_i - \mu}{\sigma}$. Then $P\underline{Z} = P \frac{\underline{X}}{\sigma} = \frac{\underline{X} - \bar{X}}{\sigma}$. Remark that P is symmetric and because it is a projection $P^2 = P$. Hence also ${}^tPP = P$ so $\sum \left(\frac{X_i - \bar{X}}{\sigma} \right)^2 = {}^t\underline{Z} {}^tP \cdot P\underline{Z} = {}^t\underline{Z} \cdot P \cdot \underline{Z} \sim \chi^2(n-1)$.

We have $\frac{s^2}{\sigma^2} \sim \frac{\chi^2(n-1)}{n-1}$. The test statistic $T = \frac{\bar{X} - \mu}{\frac{1}{\sqrt{n}} \sqrt{\frac{1}{n-1} \sum (X_i - \bar{X})^2}} =$

$$\frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{1}{n-1} \frac{\sum (X_i - \bar{X})^2}{\sigma^2}}} = \frac{\frac{\bar{X} - \mu}{\frac{\sigma}{\sqrt{n}}}}{\sqrt{\frac{s^2}{\sigma^2}}} \sim \frac{N(0,1)}{\sqrt{\frac{\chi^2(n-1)}{n-1}}} \sim t(n-1)$$

Exercise 1.6 Watch Video4 and apply a t-test to the data set Data2 to test the hypotheses

1. The data are samples from a $N(0, \sigma^2)$ distribution with σ^2 unknown
2. The data are samples from a $N(.03, \sigma^2)$ distribution with σ^2 unknown
3. The data are samples from a $N(.1, \sigma^2)$ distribution with σ^2 unknown

1.3 Analysis of Variance (ANOVA)

We now return to our original problem: testing whether the average monthly returns of INTC are constant from year to year over an 18 year time span.

Let X_{ij} denote the random variable, the return in month j of year i , so $j = 1, 2, \dots, 12$ and $i = 1, 2, \dots, 18$ where $i = 1$ denotes the year 1990 etc. We assume the following model: The X_{ij} 's are independent and

$$X_{ij} = \mu + \tau_i + \varepsilon_{ij}$$

where $\mu + \tau_i$ is the mean of the monthly returns in year i and where $\varepsilon_{ij} \sim N(0, \sigma^2)$. We assume that $\sum_i \tau_i = 0$. We should think of μ as the average over years of the average monthly return for each year.

The null-hypothesis is

$$H_0 : \tau_1 = \tau_2 = \dots = \tau_{18} = 0$$

The alternate hypothesis is

$$H_1 : \text{at least one of the } \tau_i \neq 0$$

Let \bar{X}_i denote the average in year i i.e. $\bar{X}_i = \frac{1}{n_i}(X_{i1} + X_{i2} + \dots + X_{in_i})$ and \bar{X} the average of the \bar{X}_i 's. The $mean(\bar{X}_i) = \mu + \tau_i$ and $mean(\bar{X}) = \mu$ because of the condition $\sum_i \tau_i = 0$. Consider the random variable $\sum_i n_i (\bar{X}_i - \bar{X})^2$. The mean is $\sum_i n_i \mathbb{E}(\bar{X}_i - \bar{X})^2 = \sum_i n_i (\mathbb{E}(\bar{X}_i^2) + \mathbb{E}(\bar{X}^2) - 2\mathbb{E}(\bar{X}_i \bar{X}))$.

Remark 1.7 $\mathbb{E}(X - \mathbb{E}(X))^2 = \mathbb{E}(X^2) + \mathbb{E}(X)^2 - 2\mathbb{E}(X\mathbb{E}(X)) = \mathbb{E}(X^2) + \mathbb{E}(X)^2 - 2\mathbb{E}(X)\mathbb{E}(X) = \mathbb{E}(X^2) - \mathbb{E}(X)^2$ for any random variable X

Now $\mathbb{E}(\bar{X}_i^2) = var(\bar{X}_i) + \mathbb{E}(\bar{X}_i)^2 = \sigma^2/n_i + (\mu + \tau_i)^2$ Similarly $\mathbb{E}(\bar{X}^2) = \sigma^2/kn_i + \mu^2$.

To compute $\mathbb{E}(\bar{X}_i \bar{X})$, we have $cov(\bar{X}_i \bar{X}) = \mathbb{E}((\bar{X}_i - (\mu + \tau_i))(\bar{X} - \mu)) = \mathbb{E}(\bar{X}_i \bar{X}) - (\mu + \tau_i)\mathbb{E}(\bar{X}) - \mu\mathbb{E}(\bar{X}_i) + \mu(\mu + \tau_i) = \mathbb{E}(\bar{X}_i \bar{X}) - (\mu + \tau_i)\mu - \mu(\mu + \tau_i) + \mu(\mu + \tau_i) = \mathbb{E}(\bar{X}_i \bar{X}) - \mu(\mu + \tau_i)$. Now $cov(\bar{X}_i, \bar{X}) = \frac{1}{n_i} \frac{1}{k} cov(\sum_j X_{ij}, \sum_r \bar{X}_r) =$

$\frac{1}{kn_i} \sum_j \sum_r \frac{1}{n_r} \sum_t cov(X_{ij}, X_{rt})$. Since the X_{ij} 's are assumed to be independent we have $cov(X_{ij}, X_{rt}) = 0$ unless $(i, j) = (r, t)$ and so we get $cov(\sum_j X_{ij}, \sum_r \bar{X}_r) = \sum_j \frac{1}{n_i} cov(X_{ij}, X_{ij}) = \frac{1}{n_i} \sum_j var(X_{ij}) = \sigma^2$. Hence $\mathbb{E}(\bar{X}_i \bar{X}) = \frac{1}{kn_i} \sigma^2 + \mu(\mu + \tau_i)$.

It follows that $n_i \mathbb{E}(\bar{X}_i - \bar{X})^2 = \sigma^2 - \sigma^2/k + n_i \tau_i^2$. Hence $\mathbb{E}(\sum_i n_i (\bar{X}_i - \bar{X})^2) = (k-1)\sigma^2 + \sum_i n_i \tau_i^2$. Thus $\frac{1}{k-1} \sum_i n_i (\bar{X}_i - \bar{X})^2$ is an unbiased estimator of $\sigma^2 + \sum_i n_i \tau_i^2$.

Next consider $\sum_{i,j} (X_{ij} - \bar{X}_i)^2$. The mean is $\sum_{i,j} \mathbb{E}(X_{ij} - \bar{X}_i)^2$. We have $\mathbb{E}(X_{ij} - \bar{X}_i)^2 = \mathbb{E}(X_{ij}^2) + \mathbb{E}(\bar{X}_i^2) - 2\mathbb{E}(X_{ij}\bar{X}_i)$. As above we have $\mathbb{E}(X_{ij}^2) = \sigma^2 + (\mu + \tau_i)^2$ and $\mathbb{E}(\bar{X}_i^2) = \sigma^2/n_i + (\mu + \tau_i)^2$. To compute $\mathbb{E}(X_{ij}\bar{X}_i)$ we again compute the covariance: $\text{cov}(X_{ij}, \bar{X}_i) = \mathbb{E}((X_{ij} - (\mu + \tau_i))(\bar{X}_i - (\mu + \tau_i))) = \mathbb{E}(X_{ij}\bar{X}_i) - (\mu + \tau_i)\mathbb{E}(X_{ij}) - (\mu + \tau_i)\mathbb{E}(\bar{X}_i) + (\mu + \tau_i)^2 = \mathbb{E}(X_{ij}\bar{X}_i) - (\mu + \tau_i)^2$.
 $\text{cov}(X_{ij}, \bar{X}_i) = \frac{1}{n_i} \sum_t \text{cov}(X_{ij}, X_{it})$. Again because the X_{ij} 's are independent all the covariances except $\text{cov}(X_{ij}, X_{ij}) = \text{var}(X_{ij}) = \sigma^2$ vanish. It follows that $\mathbb{E}(X_{ij}\bar{X}_i) = \sigma^2/n_i + (\mu + \tau_i)^2$ and so $\mathbb{E}(X_{ij} - \bar{X}_i)^2 = \sigma^2 - \sigma^2/n_i$. Summing over all (i, j) we get $N\sigma^2 - \sum_j \sum_i \sigma^2/n_i = N\sigma^2 - \sum_j n_i \sigma^2/n_i = (N-k)\sigma^2$.

Thus $\frac{1}{N-k} \sum_{i,j} (X_{ij} - \bar{X}_i)^2$ is an unbiased estimator of σ^2 .

The idea is that the τ_i 's are all 0 if the two estimators are not significantly different i.e. if their quotient is not significantly different from 1. In order to test this we need the distribution of this quotient.

We shall make the simplifying assumption that the n_i 's are all the same, which is certainly true in our case.

Under H_0 the vector of random variables $\underline{Z} = \begin{pmatrix} \frac{\sqrt{n}(\bar{X}_1 - \mu)}{\sigma} \\ \frac{\sqrt{n}(\bar{X}_2 - \mu)}{\sigma} \\ \vdots \\ \frac{\sqrt{n}(\bar{X}_k - \mu)}{\sigma} \end{pmatrix}$ is a

vector of standard normals, hence if we apply the orthogonal projection in the direction of $(1, 1, \dots, 1) \in \mathbb{R}^k$ we have ${}^t \underline{Z} \cdot {}^t P \cdot P \cdot \underline{Z} \sim \chi^2(k-1)$. It

is easy to verify that $P \cdot \underline{Z} = \begin{pmatrix} \sqrt{n}(\bar{X}_1 - \bar{X}) \\ \frac{\sigma}{\sqrt{n}(\bar{X}_2 - \bar{X})} \\ \vdots \\ \frac{\sigma}{\sqrt{n}(\bar{X}_k - \bar{X})} \end{pmatrix}$. Thus ${}^t\underline{Z} \cdot {}^tP \cdot P \cdot \underline{Z} = \frac{1}{\sigma^2} \sum_i n(\bar{X}_j - \bar{X})^2 \sim \chi^2(k-1)$

Consider the vector of standard normals $\underline{Z}_i = \begin{pmatrix} \frac{X_{i1} - \mu}{\sigma} \\ \frac{X_{i2} - \mu}{\sigma} \\ \vdots \\ \frac{X_{in} - \mu}{\sigma} \end{pmatrix}$. Again the

orthogonal projection in \mathbb{R}^n takes \underline{Z}_i to $\begin{pmatrix} \frac{X_{i1} - \bar{X}_i}{\sigma} \\ \frac{X_{i2} - \bar{X}_i}{\sigma} \\ \vdots \\ \frac{X_{in} - \bar{X}_i}{\sigma} \end{pmatrix}$ and so $\frac{1}{\sigma^2} \sum_j (X_{ij} - \bar{X}_i)^2 \sim$

$\chi^2(n-1)$. Summing over i we get $\frac{1}{\sigma^2} \sum_i \sum_j (X_{ij} - \bar{X}_i)^2 \sim \chi^2(n-1) + \dots +$

$\chi^2(n-1) = \chi^2(kn-k) = \chi^2(N-k)$. Thus the quotient of the two estimators

follows a $\frac{\chi^2(k-1)}{\frac{\chi^2(N-k)}{N-k}} \sim F(k-1, N-k)$ distribution which we can then use

to test the null-hypothesis.

Exercise 1.7 Watch Video 5 and perform an ANOVA test on the MSFT data

The ANOVA test relies on a number of assumptions about the model, namely all the X_{ij} need to be normal and have a common variance. To test

for common variance we can use a *Bartlett test*: this test uses the estimator

$$B = \frac{(N - k) \log(S^2) - \sum_i (n_i - 1) \log(S_i^2)}{1 + \frac{1}{3(k-1)} \left(\sum_i \frac{1}{n_i - 1} - \frac{1}{N - k} \right)}$$

where $S_i^2 = \frac{1}{n_i - 1} \sum_j (X_{ij} - \bar{X}_i)^2$ and $S^2 = \frac{1}{N - k} \sum_i (n_i - 1) S_i^2$.

It can be shown that B is approximately $\chi^2(k - 1)$ distributed and so we can use this to test the null-hypothesis that the variances are equal across years.

Bartlett's test is very sensitive to the normality assumption, another test for equality of variances that is less sensitive to this assumption is the Levene test (tests that are less sensitive to assumptions about the underlying distributions are said to be more *robust*). This test uses the test statistic

$$W = \frac{(N - k) \sum_i n_i (\bar{Z}_i - \bar{\bar{Z}})^2}{(k - 1) \sum_{i,j} (Z_{ij} - \bar{Z}_i)^2}$$

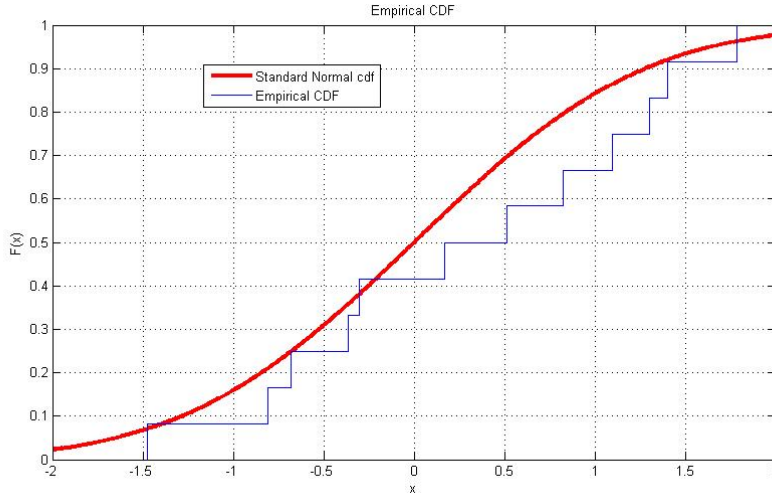
where $Z_{ij} = |X_{ij} - \bar{X}_i|$, $\bar{\bar{Z}} = \frac{1}{N} \sum_{i,j} Z_{ij}$ and $\bar{Z}_i = \frac{1}{n_i} \sum_j Z_{ij}$. The test statistic W is approximately $F(k - 1, N - k)$ distributed

Exercise 1.8 Watch Video 6 and perform Bartlett and Levene tests for equality of variances in the MSFT data

There are also tests for the data following specific distributions. An important test is the Kolmogorov-Smirnov test for normality. To describe this test we use the *empirical cdf*: let x_1, x_2, \dots, x_n be the data set but ordered so $x_1 \leq x_2 \leq \dots \leq x_n$. The empirical cdf is the piecewise constant function defined by $CDF(x_i) = i/n$ and interpolated to be constant between the x_i 's. Our null-hypothesis is:

H_0 : The data comes from a $N(\mu, \sigma^2)$ distribution

Remark that we have to specify μ and σ ahead of time, the test is not accurate if we use the sample mean and variance as parameters in the normal distribution.



The test statistic is $\max |F(x_i) - CDF(x_i)|$ where F is the cdf of the $N(\mu, \sigma^2)$ distribution. The distribution of the test statistic is tabulated and built into MATLAB.

If we do not know the μ and σ^2 parameters we can use another test for normality, the Lilliefors test, the test statistic is the same as for the Kolmogorov-Smirnov test but we are allowed to estimate the parameters from the data. Again the distribution for the test statistic (the Lilliefors distribution) is only known numerically and tabulated. The Lilliefors test is quite weak and will not reject H_0 even if the distribution of the sample data deviates from normality.

Finally the Jarque-Bera test for normality compares the higher moments, skewness and kurtosis to that of the normal distribution

Definition 1.3.1 *The n 'th (central) moment, m_n , of a random variable X is defined by $\mathbb{E}(X - \mathbb{E}(X))^n$. Thus the 2nd moment is the variance. $S = m_3/\sigma^3$ is known as the skewness and $K = m_4/\sigma^4$ as the kurtosis*

Definition 1.3.2 *The (central) moment generating function is defined by*

$$m(t) = \mathbb{E}(\exp((X - \mathbb{E}(X))t)) = \int_{-\infty}^{\infty} \exp((s - \mu)t)\phi_X(s)ds$$

Expanding the exponential function in its Taylor series, the moment gener-

ating function has a series expansion as

$$m(t) = 1 + \frac{m_2}{2}t^2 + \frac{m_3}{3!}t^3 + \frac{m_4}{4!}t^4 + \dots$$

Proposition 1.3.1 Let $X \sim N(\mu, \sigma^2)$ then the skewness is $= 0$ and the kurtosis $= 3$. All the odd central moments vanish.

Proof: We have

$$m_3 = \mathbb{E}(X - \mathbb{E}(X))^3 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} (t - \mu)^3 \exp\left(-\frac{(t - \mu)^2}{2\sigma^2}\right) dt$$

Changing variables $s = t - \mu$ the integral becomes

$$\int_{-\infty}^{\infty} s^3 \exp\left(-\frac{s^2}{2\sigma^2}\right) ds$$

Since the exponential is an even function and s^3 is odd this integral vanishes. The same argument shows that all the odd moments vanish as well. Also we see that this applies to any distribution where the density function is symmetric.

To compute the kurtosis we have

$$\begin{aligned} \sigma^2 &= \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} s^2 \exp\left(-\frac{s^2}{2\sigma^2}\right) ds \\ &= \frac{1}{\sigma\sqrt{2\pi}} \left[\frac{s^3}{3} \exp\left(-\frac{s^2}{2\sigma^2}\right) \Big|_{-\infty}^{\infty} - \int_{-\infty}^{\infty} \frac{s^3}{3} \exp\left(-\frac{s^2}{2\sigma^2}\right) \frac{-s}{\sigma^2} ds \right] \end{aligned}$$

using integration by parts. The first term on the right hand side vanishes hence we get

$$3\sigma^4 = \frac{1}{\sigma\sqrt{2\pi}} \int_{-\infty}^{\infty} s^4 \exp\left(-\frac{s^2}{2\sigma^2}\right) ds$$

The Jarque-Bera statistic is $JB = \frac{n}{6} \left(S^2 + \frac{(K - 3)^2}{4} \right)$, here $S = \frac{\frac{1}{n} \sum (X_i - \bar{X})^3}{\left(\frac{1}{n} \sum (X_i - \bar{X})^2 \right)^{3/2}}$

and $K = \frac{\frac{1}{n} \sum (X_i - \bar{X})^4}{\left(\frac{1}{n} \sum (X_i - \bar{X})^2 \right)^2}$ are estimators for the skewness and kurtosis resp.

The JB statistics is asymptotically $\chi^2(2)$ distributed.

Exercise 1.9 Watch Video 7 and perform the three tests above on the MSFT data

Our explorations so far suggests that the assumption of normality of the returns is not rejected by the data, but the assumption of constant variance across years is rejected and so the premise for the ANOVA analysis is not satisfied and the test is invalid i.e. we cannot test the null-hypothesis that the average monthly return is constant across years.

1.4 Maximum Likelihood Estimators

Let us consider a problem from the insurance industry: we are looking at the connection between getting a citation for moving violation and the probability of having an accident. When you get a traffic ticket for a moving violation your auto insurance normally goes up, but how should we quantify that.

We look at a population of drivers. We define two variables X and Y on this population. $X(i) = 1$ if individual i has had a moving violation, say within the last 3 years and 0 otherwise. $Y(i) = 1$ if the individual has had an accident within the same period and 0 if not. Thus the possible values for (X, Y) are $\{(0, 0), (0, 1), (1, 0), (1, 1)\}$. Let $P(X = 1) = p$ and $P(Y = 1) = q$. We are interested in the conditional probability $r = P(Y = 1|X = 1) = \frac{P(Y = 1, X = 1)}{P(X = 1)}$. We get $P(X = 1, Y = 1) = rp$. Since $P(X = 1) = P(X = 1, Y = 1) + P(X = 1, Y = 0)$ and $P(Y = 1) = P(X = 1, Y = 1) + P(X = 0, Y = 1)$ we get $P(X = 1, Y = 0) = p - rp$, $P(X = 0, Y = 1) = q - rp$ and since the probabilities must add up to 1 we get $P(X = 0, Y = 0) = 1 - (p - rp) - (q - rp) - rp = 1 - p - q + rp$. We can then compute the probability that from a population of n individuals we would record the values $s_{00}, s_{01}, s_{10}, s_{11}$ i.e. s_{00} is the number of individuals with no traffic violations and no accidents, s_{01} the number of individuals with no citations and at least one argument etc. The probability of this outcome is then $P = (1 - p - q + rp)^{s_{00}}(q - rp)^{s_{01}}(p - rp)^{s_{10}}(rp)^{s_{11}}$. The idea of maximum likelihood is, for a given observation, compute the values for p, q, r that maximizes P .

The values that maximize P are the same as those that maximize $\log P$ and we can simplify this maximization problem by taking log:

$$\log P = s_{00} \log(1 - p - q + rp) + s_{01} \log(q - rp) + s_{10} \log(p - rp) + s_{11} \log(rp)$$

We find the maximum by taking the partial derivatives and putting them equal to 0:

$$\frac{\partial \log P}{\partial p} = -(1-r) \frac{s_{00}}{1-p-q+rp} - r \frac{s_{01}}{q-rp} + (1-r) \frac{s_{10}}{p-rp} + r \frac{s_{11}}{rp} = 0$$

$$\frac{\log P}{\partial q} = -\frac{s_{00}}{1-p-q-rp} + \frac{s_{01}}{q-rp} = 0$$

$$\frac{\log P}{\partial r} = p \frac{s_{00}}{1-p-q+rp} - p \frac{s_{01}}{q-rp} - p \frac{s_{10}}{p-rp} + p \frac{s_{11}}{rp} = 0$$

Assume we from a population of 500 drivers, we find $s_{00} = 87, s_{01} = 152, s_{10} = 107, s_{11} = 154$. We can use the MATLAB `optimtool` to find the maximum or we can solve the equations above (watch Video 8 to see the demonstration of the `optimtool`). We find the following ML estimates: $p = .522, q = .612, r = .59$. If X and Y were independent we would have $P(Y = 1|X = 1) = P(Y = 1)$. Here we have found $P(Y = 1|X = 1) = .59$ and $P(Y = 1) = .612$, to check whether the difference is statistically significant we would have to perform a statistical test, but at least just from "eye-balling" the numbers it seems like the probability of having an accident actually goes down after being cited for a moving violation.

Consider another problem: assume we have data $\{x_1, x_2, \dots, x_n\}$ and we suspect these are sample values of independent random variables $X_i \sim N(\mu, \sigma^2)$ but we don't know μ and σ . We can of course use our previous estimators but we can use Maximum Likelihood as follows: the joint density of (X_1, X_2, \dots, X_n) is $\phi_{X_1, X_2, \dots, X_n}(t_1, t_2, \dots, t_n | \mu, \sigma) = \frac{1}{\sigma^n (2\pi)^{n/2}} \prod \exp(-\frac{(t_i - \mu)^2}{2\sigma^2})$. If $\Delta t_1, \dots, \Delta t_n$ are small we can interpret $\phi_{X_1, X_2, \dots, X_n}(t_1, t_2, \dots, t_n | \mu, \sigma) \Delta t_1 \Delta t_2 \dots \Delta t_n$ as the probability that a sample value (x_1, x_2, \dots, x_n) lies in the n -dimensional box with vertex at (t_1, t_2, \dots, t_n) and sides $\Delta t_1, \Delta t_2, \dots, \Delta t_n$. Thus we want to find μ and σ such that $\phi_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n | \mu, \sigma) \Delta x_1 \Delta x_2 \dots \Delta x_n$ is maximal for our given samples $\{x_1, x_2, \dots, x_n\}$. If the Δx 's are fixed this means finding μ and σ such that $\phi_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n | \mu, \sigma)$ is maximal. The function $\mathcal{L}(\mu, \sigma) = \phi_{X_1, X_2, \dots, X_n}(x_1, x_2, \dots, x_n | \mu, \sigma)$ is known as the *likelihood function*. Finding maximum of this function is equivalent to finding maximum of the *log-likelihood function* $\ell(\mu, \sigma) = \log \mathcal{L}(\mu, \sigma) = -n \log \sigma - (n/2) \log 2\pi + \sum -\frac{(x_i - \mu)^2}{2\sigma^2}$.

The partial derivatives are given by $\frac{\partial}{\partial \mu} \ell(\mu, \sigma) = \sum \frac{x_i - \mu}{2\sigma^2}$ and $\frac{\partial}{\partial \sigma} \ell(\mu, \sigma) = -\frac{n}{\sigma} + \sum \frac{(x_i - \mu)^2}{\sigma^3}$. Setting these equal to 0 gives $\mu = \frac{1}{n} \sum x_i = \bar{x}$ and $\sigma^2 = \frac{1}{n} \sum (x_i - \bar{x})^2$. Remark that the maximum likelihood estimate for σ is biased.

Here is yet another example: consider a linear model $Y_i = a + bX_i + \varepsilon_i$ where $\varepsilon_i \sim N(0, \sigma^2)$. Suppose we have data points $(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)$. We want to find the maximum likelihood estimates for the coefficients a and b . Given x_i and σ the distribution of Y_i is $N(a + bx_i, \sigma^2)$ so the joint distribution of Y_1, Y_2, \dots, Y_n given x_1, x_2, \dots, x_n and σ is $\phi_{Y_1, \dots, Y_n}(t_1, t_2, \dots, t_n | x_1, x_2, \dots, x_n, \sigma) = \frac{1}{\sigma^n (2\pi)^{n/2}} \prod \exp\left(-\frac{(t_i - (a + bx_i))^2}{2\sigma^2}\right)$. It follows that the likelihood function is $\mathcal{L}(a, b, \sigma) = \frac{1}{\sigma^n (2\pi)^{n/2}} \prod \exp\left(-\frac{(y_i - (a + bx_i))^2}{2\sigma^2}\right)$ and the log-likelihood function $\ell(a, b, \sigma) = -n \log \sigma - (n/2) \log 2\pi + \sum -\frac{(y_i - (a + bx_i))^2}{2\sigma^2}$. Taking partials w.r.t. a and b and setting them equal to 0 we get the maximum likelihood estimates $\hat{a} = \frac{\sum y_i \sum x_i^2 - \sum y_i x_i \sum x_i}{n \sum x_i^2 - (\sum x_i)^2}$ and $\hat{b} = \frac{\sum y_i \sum x_i - n \sum y_i x_i}{(\sum x_i)^2 - n \sum x_i^2}$. These are precisely the *least squares estimates*.