The t Distributions

Low-tech Derivation

If $X_1, \ldots, X_n$ are i.i.d. with mean m and sample average

$$A := \frac{1}{n} \sum_{i=1}^{n} X_i \,,$$

then, if we don't know m, we use A to estimate m. How far off is this likely to be? If we don't know the SD, σ, of X_i, then we use the modified sample variance

$$V^+ := \frac{1}{n-1} \sum_{i=1}^{n} (X_i - A)^2$$

to estimate the SE of A as $\sqrt{V^+/n}$ instead of as $\sigma/\sqrt{n}$. We know that the normalized error $z := (A - m)/(\sigma/\sqrt{n})$ has approximately a standard normal distribution, written $\mathrm{N}(0, 1)$. This is exact if X_i are themselves normally distributed because we are just standardizing A (i.e., subtracting the mean of A and then dividing by the SD of A) and A is itself a sum of independent normal random variables, whence normal.

But what is the distribution of $t := (A - m)/\sqrt{V^+/n}$? Now we are dividing by a random variable, not by a constant. For very large n, the denominator is almost constant, so t is almost the same as z and therefore t is almost normal. But for small samples, this is not the case, as explained in Chap. 26, Section 6, of FPP. Here, we'll give a simple expression for t in terms of normal random variables when X_i are themselves normal, but we won't go as far as getting the density of t.

Note that we will now assume that X_i are normal. First, this does not mean that the distribution of t is normal; we will calculate what it is. Second, although the figures on p. 491 of FPP show that the distributions of t and z look close to each other, the differences are important because they matter most in the tails, where one may be several times the other. This may easily make the difference between accepting or rejecting the null hypothesis (and thus to getting a paper published or not!). Third, it does not mean that if X_i are not normal, then the distribution of t is approximately the same as if X_i were normal. This is a distinction from z. For more on this distinction, see note 10 on p. A-29 of FPP. For another example, I simulated 20 samples from an exponential distribution with mean 1 and computed t. I did this 10,000 times. If the distribution were normal instead of exponential, then there would be a value $t_{0.05}$ such that $P\big(|t| > t_{0.05}\big) = 0.05$; this value

is the 5% significance level of t. (In this case, $t_{0.05} \approx 2.09$ by the table on p. A-105 of FPP.) Thus, the number of times that $|t|$ would turn out to be larger than $t_{0.05}$ would be about 500 in 10,000 simulations from a normal distribution. In actuality, however, with the exponential distribution, it turned out that large 872 times. The difference (between 500 and 872) is too large to account for by sampling error: the SE for the sum of 10,000 draws from a box with 5% ones is only 22.

Note that if we standardize X_i, that will not change t, so ***we will assume from the start that*** X_i ***are*** $\mathrm{N}(0,1)$ ***instead of*** $\mathrm{N}(m,\sigma^2)$**.**

Suppose that $(w_1,\dots,w_n)$ is an orthonormal basis of $\mathbb{R}^n$. Later, we will choose a particularly useful one. This means that $w_i \cdot w_j = \delta_{i,j}$, where

$$\delta_{i,j} := \begin{cases} 0 & \text{if } i \neq j, \\ 1 & \text{if } i = j. \end{cases}$$

Note that we also have $E(X_iX_j) = \delta_{i,j}$. Write the coordinates of w_j as $(w_{j,1},\dots,w_{j,n})$. Define $Y_j := \sum_{i=1}^n X_i w_{j,i}$. (Later, we'll explain where all this comes from.) By our definition of multi-variate normal distribution, we know that $(Y_1,\dots,Y_n)$ has a multi-variate normal distribution. But which one? We know it suffices to determine the means and covariances. Now

$$E(Y_j) = \sum_{i=1}^n E(X_i w_{j,i}) = \sum_{i=1}^n E(X_i) w_{j,i} = 0$$

and

$$\begin{aligned} E(Y_jY_p) &= E\big(\sum_{i=1}^n X_i w_{j,i} \sum_{k=1}^n X_k w_{p,k}\big) = E\big(\sum_{i=1}^n \sum_{k=1}^n X_i w_{j,i} X_k w_{p,k}\big) \\ &= \sum_{i=1}^n \sum_{k=1}^n E(X_i w_{j,i} X_k w_{p,k}) = \sum_{i=1}^n \sum_{k=1}^n E(X_iX_k) w_{p,k} w_{j,i} \\ &= \sum_{i=1}^n \sum_{k=1}^n \delta_{i,k} w_{p,k} w_{j,i} = \sum_{i=1}^n w_{p,i} w_{j,i} = w_p \cdot w_j \\ &= \delta_{p,j}\,. \end{aligned}$$

In other words, the means and covariances of Y_j are the same as those of X_i, so we know this implies they have the same distribution, i.e., they are are i.i.d. $\mathrm{N}(0,1)$. (This looks very surprising here, but we'll understand this better later.)

In order to perform another miracle, we need to recall the following fact from linear algebra. Write $R := [\, w_1 \quad w_2 \quad \cdots \quad w_n \,]$ for the matrix whose columns are the column vectors $w_1, w_2, \dots, w_n$. Write R' for its transpose and I_n for the $n \times n$ identity matrix. Then

the statement that $(w_1, \dots, w_n)$ is an orthonormal basis is equivalent to the statement that $R'R = I_n$. (Remember that to multiply matrices, we multiply rows from the left matrix by columns from the right matrix, which is the same as taking dot products.) But this implies that R' is the inverse of R, so we also have $RR' = I_n$. If we write out this last matrix multiplication, we get that

$$\sum_{j=1}^{n} w_{j,i} w_{j,k} = \delta_{i,k} \,.$$

Now look at this:

$$\begin{aligned}
\sum_{j=1}^{n} Y_j^2 &= \sum_{j=1}^{n} \big(\sum_{i=1}^{n} X_i w_{j,i}\big)^2 = \sum_{j=1}^{n}\sum_{i=1}^{n} X_i w_{j,i} \sum_{k=1}^{n} X_k w_{j,k} \\
&= \sum_{j=1}^{n}\sum_{i=1}^{n}\sum_{k=1}^{n} X_i w_{j,i} X_k w_{j,k} = \sum_{i=1}^{n}\sum_{k=1}^{n}\sum_{j=1}^{n} X_i X_k w_{j,i} w_{j,k} \\
&= \sum_{i=1}^{n}\sum_{k=1}^{n} X_i X_k \sum_{j=1}^{n} w_{j,i} w_{j,k} = \sum_{i=1}^{n}\sum_{k=1}^{n} X_i X_k \delta_{i,k} \\
&= \sum_{i=1}^{n} X_i^2 \,.
\end{aligned}$$

Wow!

With these miracles in hand, we are ready to analyze the t-distribution. Suppose that we take our orthonormal basis so that it starts with $w_1 := (1, 1, \dots, 1)/\sqrt{n}$. It doesn't matter what the other w_j are. Then $Y_1 = \sum_{i=1}^{n} X_i w_{1,i} = \sum_{i=1}^{n} X_i/\sqrt{n} = \sqrt{n}A$, i.e., $A = Y_1/\sqrt{n}$. Furthermore, we can use the alternative formula for variance and the preceding miracle to write

$$\begin{aligned}
V &:= \frac{1}{n}\sum_{i=1}^{n}(X_i - A)^2 = \frac{1}{n}\sum_{i=1}^{n} X_i^2 - A^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - A^2 = \frac{1}{n}\sum_{i=1}^{n} Y_i^2 - \frac{1}{n} Y_1^2 \\
&= \frac{1}{n}\sum_{i=2}^{n} Y_i^2 \,,
\end{aligned}$$

so that

$$V^+ = \frac{1}{n-1}\sum_{i=2}^{n} Y_i^2 \,.$$

Since Y_i are all independent, this shows that V^+ is independent of Y_1 and so of A. Now $t = A/\sqrt{V^+/n}$ because we are taking $m = 0$. Therefore, we have proved that the numerator

and denominator of t are independent! (In words, the sample average and the sample variance are independent!) If we use the equation $A = Y_1/\sqrt{n}$, we can cancel the factors of $\sqrt{n}$ to get $t = Y_1/\sqrt{V^+}$, i.e.,

$$t = \frac{Y_1}{\sqrt{\frac{1}{n-1}\sum_{i=2}^n Y_i^2}},$$

where Y_i are i.i.d. $\mathrm{N}(0,1)$. This is our final result. By the way, the sum of squares of d i.i.d. $\mathrm{N}(0,1)$ random variables has a distribution that is called χ^2_d. (That's the Greek letter chi, pronounced "kye" to rhyme with "bye".) There is one t-distribution for each d; in this context, d is called the "degrees of freedom" and we use $d = n-1$.

High-tech Derivation

If you want to understand what's really going on here, as well as in most of SM, you need to use more linear algebra.

Given random variables $X_1, \ldots, X_n$, form the random vector $X := (X_1, \ldots, X_n) \in \mathbb{R}^n$. We'll regard this as a column vector. We'll use some manipulations of random vectors that are covered in SM, Chap. 3. To say that X_i are i.i.d. $\mathrm{N}(0,1)$ is the same as saying that X is $\mathrm{N}(\mathbf{0}_n, I_n)$, where the first entry $\mathbf{0}_n := (0, 0, \ldots, 0)$ is the mean $E(X)$ and the second entry is the covariance matrix $E(XX')$. We call this distribution a ***standard normal distribution of dimension*** n.

A matrix whose columns form an orthonormal basis is called an orthogonal matrix, as R was above. It is a change-of-basis matrix: multiplying by R' changes a coordinate vector x to the corresponding coordinates $R'x$ in the basis of the columns of R (since $x = I_n x = (RR')x = R(R'x)$, which is the linear combination of the columns of R with coefficients from $R'x$). (This should also be familiar from the way we calculate coordinates in an *orthonormal* basis: let C_i be the ith column of R. To say that the coordinates of x in the basis $(C_1, \ldots, C_n)$ are $(a_1, \ldots, a_n)$ is to say that $x = \sum_{i=1}^n a_i C_i$. Now in this case, for all j, we have $C_j \cdot x = \sum_{i=1}^n a_i C_j \cdot C_i = a_j$ since $C_j \cdot C_i = \delta_{i,j}$. Thus, the coordinates are just the dot products with the basis vectors. Furthermore, the jth coordinate of $R'x$ is equal to the dot product of the jth row of R' with x, i.e., $C_j \cdot x$.) Thus, $Y = R'X$ is the vector X in new coordinates. We proved that, if X has the distribution $\mathrm{N}(\mathbf{0}_n, I_n)$, then Y also has the distribution $\mathrm{N}(\mathbf{0}_n, I_n)$. Here's a matrix proof:

$$E(Y) = E(R'X) = R'E(X) = R'\mathbf{0}_n = \mathbf{0}_n$$

and

$$E(YY') = E(R'XX'R) = R'E(XX')R = R'I_nR = R'R = I_n\,.$$

Wasn't that nice? We also proved that $\|Y\|^2 = \|X\|^2$; here's a nice matrix proof:

$$\|Y\|^2 = Y \cdot Y = Y'Y = X'RR'X = X'I_nX = X'X = X \cdot X = \|X\|^2\,.$$

Of course, since Y is just the vector X in different orthonormal coordinates, we could also conclude this equality $\|Y\| = \|X\|$ immediately. Since the coordinates of Y are linear combinations of the coordinates of X, we know that Y has a multi-variate normal distribution, and we have just calculated its parameters. (By the way, our calculation shows that even without assuming X to be normal, as long as $E(X) = \mathbf{0}_n$, we have $E(Y) = \mathbf{0}_n$; as long as the covariance matrix of X is I_n, so is the covariance matrix of Y; and in all cases, $\|Y\| = \|X\|$.)

Another way to see this—for standard normal distributions—is by using the density of X. Multiplication by an orthogonal matrix is, geometrically, a rotation of the vector it multiplies (combined possibly with a change of sign of some coordinate). The standard normal n-dimensional distribution $\mathrm{N}(\mathbf{0}_n, I_n)$ is rotationally symmetric about the origin, so it doesn't change under rotations. That's why Y and X have the same distribution. (We have now proved this in two ways. To see this symmetry a third way with the density, use the fact that X_i has density $e^{-x^2/2}/\sqrt{2\pi}$, so that X has density

$$\prod_{i=1}^{n} e^{-x_i^2/2}/\sqrt{2\pi} = e^{\sum_{i=1}^{n} -x_i^2/2}/(2\pi)^{n/2} = e^{-\|x\|^2/2}/(2\pi)^{n/2}\,.$$

Since this depends only on $\|x\|$, we see the symmetry.) Also, R does not change the length of any vector, so X and Y have the same length.

(Recall, though, that we proved, even without assuming X to be normal, that as long as $E(X) = \mathbf{0}_n$, we have $E(Y) = \mathbf{0}_n$; as long as the covariance matrix of X is I_n, so is the covariance matrix of Y; and in all cases, $\|Y\| = \|X\|$.)

We started with a random vector X whose distribution was $\mathrm{N}(\mathbf{0}_n, I_n)$ and obtained a random vector $Y = R'X$ with the same distribution. In other words, the coordinates of Y are independent $\mathrm{N}(0,1)$ random variables, just like the coordinates of X are. Recall that Y is just X in different orthonormal coordinates. This is convenient for taking orthogonal projections of X, since we can choose a convenient orthonormal basis.

Thus, suppose that W is a subspace of $\mathbb{R}^n$ of dimension d. Choose an orthonormal basis $(w_1, \ldots, w_d)$ of W and an orthonormal basis $(w_{d+1}, \ldots, w_n)$ of $W^\perp$. Then $(w_1, \ldots, w_n)$

is an orthonormal basis of $\mathbb{R}^n$. Let $R := [\, w_1 \quad w_2 \quad \cdots \quad w_n \,]$ be the corresponding orthogonal matrix. We saw that the coordinates of X in the basis $(w_1, \ldots, w_n)$ are the coordinates of $R'X = Y = (Y_1, Y_2, \ldots, Y_d, Y_{d+1}, \ldots, Y_n)$. In other words, $X = \sum_{i=1}^n Y_i w_i$. Thus, the coordinates of $P_W(X)$ in the basis $(w_1, \ldots, w_d)$ are $(Y_1, \ldots, Y_d)$. Since Y has a $\mathrm{N}(\mathbf{0}_n, I_n)$ distribution, we know that $(Y_1, \ldots, Y_d)$ has a $\mathrm{N}(\mathbf{0}_d, I_d)$ distribution. Likewise, $(Y_{d+1}, \ldots, Y_n)$ has a $\mathrm{N}(\mathbf{0}_{n-d}, I_{n-d})$ distribution. Finally, these two random vectors, $(Y_1, \ldots, Y_d)$ and $(Y_{d+1}, \ldots, Y_n)$ are independent of each other (since *all* the coordinates of Y are independent). In other words, $P_W(X)$ and $P_{W^\perp}(X)$ are both independent standard normal random vectors, the first being of dimension d and the second of dimension $n - d$. Here, we are viewing $P_W(X)$ as a d-dimensional vector, i.e., as a vector in the vector space W, rather than as an n-dimensional vector that happens to lie in W; likewise for $P_{W^\perp}(X)$.

Now to go back to the t-distribution. We take W to be the span of $(1, 1, \ldots, 1)$. Then $w_1 := (1, 1, \ldots, 1)/\sqrt{n}$ is by itself an orthonormal basis for W. In particular, $Y_1 = w_1 \cdot X$. Thus, $Y_1/\sqrt{n} = A$, which is the numerator of $t = A/\sqrt{V^+/n}$. The denominator of t uses the vector of deviations from the average,

$$\begin{aligned}(X_1 - A, X_2 - A, \ldots, X_n - A) &= X - A(1, 1, \ldots, 1) = X - \frac{Y_1}{\sqrt{n}}(1, 1, \ldots, 1) \\ &= X - Y_1 w_1 = X - P_W(X) = P_{W^\perp}(X)\,.\end{aligned}$$

Thus,

$$V^+ = \frac{1}{n-1}\|P_{W^\perp}(X)\|^2 = \frac{1}{n-1}\sum_{i=2}^n Y_i^2\,,$$

which means that

$$t = \frac{Y_1}{\sqrt{V^+}} = \frac{Y_1}{\sqrt{\sum_{i=2}^n Y_i^2}/\sqrt{n-1}}\,.$$

As we saw, all Y_i have distribution $\mathrm{N}(0, 1)$ and are independent, so the numerator and denominator of t are independent.

Later, in multiple regression we will use subspaces W of dimension higher than 1. A key role will be played by understanding how random vectors behave with respect to the orthogonal projections onto W and onto $W^\perp$, as we discussed above.