

Student's t distribution

Student's t distribution is very important for the development of inference procedures about a population mean m . We all stress the fact that the t -distribution approaches the normal distribution as the sample size, n , goes to infinity, and point out that the p -values in the last row (highest degrees of freedom) of the table of t -distribution critical values are almost the same as those from the z -distribution. In this section, we will define the t -distribution, prove that $t = \frac{\bar{Y} - m}{S/\sqrt{n}}$ has a t -distribution with $n-1$ degrees of freedom, and

prove that the limiting properties we describe in class are true.

A random variable that possesses a Student's t distribution is defined as the ratio of a standard normal random variable Z divided by the square root of an independent chi-square random variable which has been divided by its degrees of freedom. Stated mathematically: let $Z \sim N(0,1)$ be independent of $W \sim \chi_n^2$. Then $T = \frac{Z}{\sqrt{W/n}}$ is said to have a t distribution with n degrees of freedom. Note that independence of Z and W is very important.

It can be shown that the density function for T is $f_T(t) = \frac{\Gamma[(n+1)/2]}{\sqrt{pn}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2}$,
 $-\infty < t < \infty$.

Theorem: $\lim_{n \rightarrow \infty} f_T(t) = \frac{1}{\sqrt{2p}} e^{-t^2/2}$, which is the density function for a standard normal random variable.

Proof: First look at $\left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} = \left\{ \left(1 + \frac{t^2}{n}\right)^{n/t^2} \right\}^{-t^2/2} \left(1 + \frac{t^2}{n}\right)^{-1/2}$

To evaluate $\lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{n}\right)^{n/t^2}$, let $h = \frac{n}{t^2}$. Then as $n \rightarrow \infty, h \rightarrow \infty$.

So $\lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{n}\right)^{n/t^2} = \lim_{h \rightarrow \infty} \left(1 + \frac{1}{h}\right)^h = e$.

Note that $\lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{n}\right)^{-1/2} = 1$.

$$\begin{aligned} \text{So } \lim_{n \rightarrow \infty} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} &= \lim_{n \rightarrow \infty} \left\{ \left(1 + \frac{t^2}{n}\right)^{n/t^2} \right\}^{-t^2/2} \left(1 + \frac{t^2}{n}\right)^{-1/2} \\ &= e^{-t^2/2} \cdot 1 = e^{-t^2/2} \end{aligned}$$

Note that $\lim_{n \rightarrow \infty} \frac{\Gamma[(n+1)/2]}{\sqrt{pn}\Gamma(n/2)} = \frac{1}{\sqrt{2p}}$. Although this is far from obvious, it can be shown using Stirling's formula (see Appendices 1 and 2). Also, we don't really need to worry about proving that $\frac{1}{\sqrt{2p}}$ this is the correct constant, since all that matters is the form of the limiting density function. We know that the constant must make the integral equal to 1, so since the density function for $N(0,1)$ is proportional to $e^{-t^2/2}$. That information alone requires the constant of proportionality to be $\frac{1}{\sqrt{2p}}$.

Putting all the pieces together, we have:

$$\lim_{n \rightarrow \infty} f_T(t) = \lim_{n \rightarrow \infty} \frac{\Gamma[(n+1)/2]}{\sqrt{pn}\Gamma(n/2)} \left(1 + \frac{t^2}{n}\right)^{-(n+1)/2} = \frac{1}{\sqrt{2p}} e^{-t^2/2}.$$

In practice, we use the t distribution for inference procedures on means when the population variance is not known. Therefore, we want to show that $\frac{\bar{Y} - \mathbf{m}}{S/\sqrt{n}}$ (where S^2 is the sample variance) has a t distribution. To prove that this is true we will need to prove three preliminary theorems.

Theorem 1: Let Y_1, Y_2, \dots, Y_n be independent identically distributed (iid) normal random variables with mean \mathbf{m} and variance \mathbf{s}^2 . Then $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ is normally distributed with mean \mathbf{m} and variance \mathbf{s}^2/n .

Proof: $m_{Y_i}(t) = e^{\mathbf{m}t + \frac{\mathbf{s}^2 t^2}{2}}$

$$m_{\frac{1}{n}Y_i}(t) = m_{Y_i}\left(\frac{1}{n}t\right) = e^{\frac{\mathbf{m}t}{n} + \frac{\mathbf{s}^2 t^2}{2n^2}}$$

Note that $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i = \frac{1}{n} Y_1 + \frac{1}{n} Y_2 + \dots + \frac{1}{n} Y_n$, and since the random variables

Y_i with $i=1,2,3,\dots,n$ are independent, the random variables $\frac{1}{n} Y_i$ with $i=1,2,3,\dots,n$ are also independent.

Therefore: $m_{\bar{Y}}(t) = m_{(1/n)Y_1}(t) \cdot m_{(1/n)Y_2}(t) \cdots m_{(1/n)Y_n}(t)$

$$= \left(e^{\frac{mt + \frac{s^2 t^2}{2n}}{n}} \right)^n = e^{\frac{mt + \frac{s^2 t^2}{2n}}{1}}, \text{ which is the moment generating}$$

function for a normal random variable with mean m , variance $\frac{s^2}{n}$.

$\therefore \bar{Y} \sim N\left(m, \frac{s^2}{n}\right)$ by the uniqueness theorem.

Independence of the Sample Mean and Sample Variance

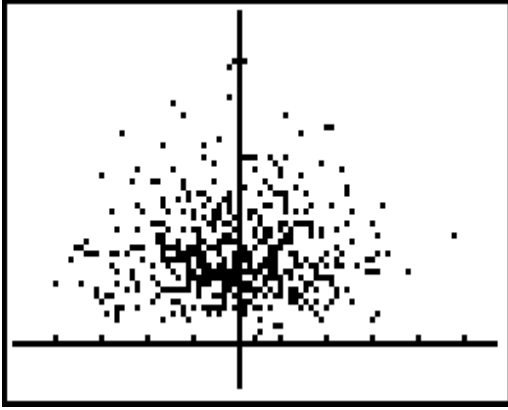
If the sample mean is large, does this give us any information about the sample variance? Next we prove a theorem that may seem surprising, the independence of the mean \bar{Y} and variance S^2 . Since S^2 uses \bar{Y} in its computation, how can they possibly be independent? First we will generate a simple simulation to illustrate this independence. Then we will develop the proof using moment generating functions.

TI-83 Program for Simulation

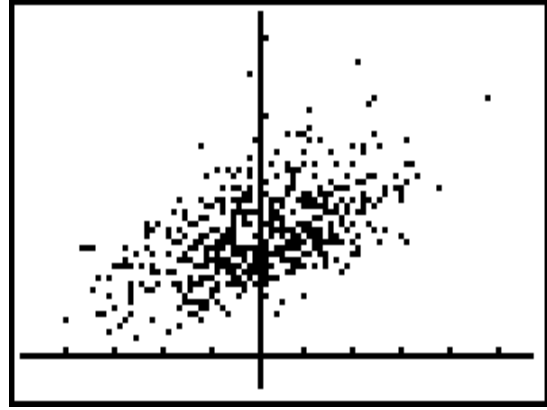
This program selects 10 random numbers from the $N(0,1)$ distribution, computes the mean, the variance, finds the maximum of the 10 numbers, and finally stores the mean in List 1, the variance in List 2, and the maximum of the 10 numbers in List 3. It repeats this process 500 times. We will then plot List 2 against List 1. If the values of the mean and variance are independent, we should see only a random scattering of ordered pairs. If they are dependent, then we should see a linear pattern in the data, as we see when we plot the maximum against the mean.

PROGRAM: INDEPEN

```
: For (I, 1, 500)
: randNorm(0,1,10) ® L6
: mean(L6) → L1(I)
: variance(L6) → L2(I)
: max(L6) → L3(I)
: End
```



Variance vs Mean



Maximum vs Mean

Figure 4: Scatterplots of Variance and Maximum against Mean

Theorem 2: Let Y_1, Y_2, \dots, Y_n be independent identically distributed normal random variables with mean μ and variance σ^2 . Then $\bar{Y} = \frac{1}{n} \sum_{i=1}^n Y_i$ and $S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$ are independent.

Lemma: \bar{Y} is independent of $(Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y})$.

Proof: $m_{\bar{Y}, Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y}}(t, t_1, \dots, t_n) = E\{\exp[t\bar{Y} + t_1(Y_1 - \bar{Y}) + \dots + t_n(Y_n - \bar{Y})]\}$

Note that $[t\bar{Y} + t_1(Y_1 - \bar{Y}) + \dots + t_n(Y_n - \bar{Y})] = t\bar{Y} + \sum_i t_i(Y_i - \bar{Y})$

$$\begin{aligned} &= t \frac{\sum_i Y_i}{n} + \sum_i (t_i Y_i - t_i \bar{Y}) = \sum_i \frac{t}{n} Y_i + \sum_i t_i Y_i - \sum_i t_i \bar{Y} \\ &= \sum_i \frac{t}{n} Y_i + \sum_i t_i Y_i - \bar{Y} \sum_i t_i = \sum_i \frac{t}{n} Y_i + \sum_i t_i Y_i - \frac{\sum_i t_i}{n} \sum_i Y_i \\ &= \sum_i \frac{t}{n} Y_i + \sum_i t_i Y_i - \sum_i Y_i \sum_i \frac{t_i}{n} = \sum_i \frac{t}{n} Y_i + \sum_i t_i Y_i - \sum_i Y_i \bar{t} \\ &= \sum_i \left[\frac{t}{n} + t_i - \bar{t} \right] Y_i = \sum_i \left[\frac{t}{n} + (t_i - \bar{t}) \right] Y_i = \sum_i a_i Y_i \end{aligned}$$

$$\text{where } a_i = \frac{t}{n} + (t_i - \bar{t})$$

Note that $\sum_i a_i Y_i$ is a linear combination of independent normal random variables.

Also
$$\sum_i a_i = \sum_i \frac{t}{n} + (t_i - \bar{t}) = n \cdot \frac{t}{n} + \sum_i (t_i - \bar{t}) = t + 0 = t,$$

and
$$\begin{aligned} \sum_i a_i^2 &= \sum_i \left[\frac{t}{n} + (t_i - \bar{t}) \right]^2 = \sum_i \left[\frac{t^2}{n^2} + 2 \frac{t}{n} (t_i - \bar{t}) + (t_i - \bar{t})^2 \right] \\ &= n \cdot \frac{t^2}{n^2} + 2 \frac{t}{n} \sum_i (t_i - \bar{t}) + \sum_i (t_i - \bar{t})^2 = \frac{t^2}{n} + \sum_i (t_i - \bar{t})^2. \end{aligned}$$

Thus, $m_{\bar{Y}, Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y}}(t, t_1, \dots, t_n) = E\{\exp[\sum a_i Y_i]\}$

$$\begin{aligned} &= \prod_{i=1}^n m_{Y_i}(a_i) = \prod_{i=1}^n \exp\left[\mathbf{m} a_i + \frac{\mathbf{s}^2}{2} a_i^2 \right] \\ &= e^{(\mathbf{m} a_1 + \mathbf{m} a_2 + \dots + \mathbf{m} a_n) + \frac{\mathbf{s}^2 (a_1^2 + a_2^2 + \dots + a_n^2)}{2}} = e^{\mathbf{m} \sum a_i + \frac{\mathbf{s}^2 \sum a_i^2}{2}} \\ &= e^{\mathbf{m} + \frac{\mathbf{s}^2}{2} \left[\frac{t^2}{n} + \sum (t_i - \bar{t})^2 \right]} = e^{\mathbf{m} + \frac{\mathbf{s}^2 t^2}{2n} + \frac{\mathbf{s}^2 \sum (t_i - \bar{t})^2}{2}} \\ &= e^{\mathbf{m} + \frac{\mathbf{s}^2 t^2}{2n}} e^{\frac{\mathbf{s}^2 \sum (t_i - \bar{t})^2}{2}} = m_{\bar{Y}}(t) \cdot h(t_1, t_2, \dots, t_n) \end{aligned}$$

Since we have the partial factorization $m_{\bar{Y}, Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y}} = m_{\bar{Y}}(t) \cdot h(t_1, t_2, \dots, t_n)$, it follows that $h(t_1, t_2, \dots, t_n)$ must be the joint moment generating function of $Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y}$, and that \bar{Y} is independent of $Y_1 - \bar{Y}, Y_2 - \bar{Y}, \dots, Y_n - \bar{Y}$. The sample mean and sample variance of a simple random sample from a normal distribution are independent.

Theorem 3: Let Y_1, Y_2, \dots, Y_n be a random sample from a normal distribution with mean \mathbf{m}

and variance \mathbf{s}^2 . Then $\frac{(n-1)S^2}{\mathbf{s}^2} \sim \mathbf{c}_{n-1}^2$ where $S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}$.

Proof:
$$\begin{aligned} \sum_{i=1}^n \left(\frac{Y_i - \mathbf{m}}{\mathbf{s}} \right)^2 &= \sum_{i=1}^n \frac{[(Y_i - \bar{Y}) + (\bar{Y} - \mathbf{m})]^2}{\mathbf{s}^2} \\ &= \frac{\sum (Y_i - \bar{Y})^2}{\mathbf{s}^2} + \frac{\sum (\bar{Y} - \mathbf{m})^2}{\mathbf{s}^2} + \frac{2}{\mathbf{s}^2} \sum (Y_i - \bar{Y})(\bar{Y} - \mathbf{m}) \\ &\quad \text{Note: } \frac{2}{\mathbf{s}^2} \sum (Y_i - \bar{Y})(\bar{Y} - \mathbf{m}) = \frac{2}{\mathbf{s}^2} (\bar{Y} - \mathbf{m}) \sum (Y_i - \bar{Y}) = 0 \\ &= \frac{\sum (Y_i - \bar{Y})^2}{\mathbf{s}^2} + \frac{\sum (\bar{Y} - \mathbf{m})^2}{\mathbf{s}^2} \\ &= \frac{(n-1)S^2}{\mathbf{s}^2} + \frac{n(\bar{Y} - \mathbf{m})^2}{\mathbf{s}^2} \end{aligned}$$

$$\therefore \sum_{i=1}^n \left(\frac{Y_i - \mathbf{m}}{\mathbf{s}} \right)^2 = \frac{(n-1)S^2}{\mathbf{s}^2} + \frac{n(\bar{Y} - \mathbf{m})^2}{\mathbf{s}^2} = \frac{(n-1)S^2}{\mathbf{s}^2} + \frac{(\bar{Y} - \mathbf{m})^2}{\left(\frac{\mathbf{s}^2}{n} \right)} \quad (1)$$

Note that $\sum_{i=1}^n \left(\frac{Y_i - \mathbf{m}}{\mathbf{s}} \right)^2$ is the sum of n Z^2 variables, which (as previously shown) has a \mathbf{c}_n^2 distribution, and that $\frac{n(\bar{Y} - \mathbf{m})^2}{\mathbf{s}^2} = \frac{(\bar{Y} - \mathbf{m})^2}{\mathbf{s}^2/n}$ is the square of a standard normal variable, which has a \mathbf{c}_1^2 distribution. Moment generating functions can be used to prove that $\mathbf{c}_a^2 + \mathbf{c}_b^2 \sim \mathbf{c}_{a+b}^2$ if the random variables are independent. This result in combination with the result in equation (1) leads to the conclusion that $\frac{(n-1)S^2}{\mathbf{s}^2} \sim \mathbf{c}_{n-1}^2$.

Using Theorems 1, 2, and 3, we can now prove what we originally set out to prove:

Theorem: Let Y_1, Y_2, \dots, Y_n be independent identically distributed random variables, each normally distributed with mean \mathbf{m} and variance \mathbf{s}^2 .

Then $\frac{\bar{Y} - \mathbf{m}}{S/\sqrt{n}} \sim t_{n-1}$.

Proof: $Z = \frac{\bar{Y} - \mathbf{m}}{\mathbf{s}/\sqrt{n}} \sim N(0,1)$ by Theorem 1.

$\frac{(n-1)S^2}{\mathbf{s}^2} \sim \mathbf{c}_{n-1}^2$ by Theorem 2.

$Z = \frac{\bar{Y} - \mathbf{m}}{\mathbf{s}/\sqrt{n}}$ is independent of $\frac{(n-1)S^2}{\mathbf{s}^2}$ by Theorem 3.

Thus,

$$T = \frac{\frac{\bar{Y} - \mathbf{m}}{\mathbf{s}/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\mathbf{s}^2(n-1)}}} \text{ is in the form of } \frac{Z}{\sqrt{\mathbf{c}_{n-1}^2/n-1}}$$

So by the definition of t ,

$$T = \frac{\frac{\bar{Y} - \mathbf{m}}{\mathbf{s}/\sqrt{n}}}{S/\mathbf{s}} = \frac{\bar{Y} - \mathbf{m}}{S/\sqrt{n}}$$

is distributed as t with $n-1$ degrees of freedom.

Note that \mathbf{s} disappears in the formula for T .

Expected Value of S^2

One question that always comes up in every Introductory Statistics class is, "Why divide by $n-1$ instead of n when find the sample standard deviation?" The answer is that the expected value of S^2 is \mathbf{s}^2 .

First we will assume normally distributed random variables and show that $E(S^2) = \mathbf{s}^2$.

Theorem: Let Y_1, Y_2, \dots, Y_n be independent identically distributed random variables, each normally distributed with mean \mathbf{m} and variance \mathbf{s}^2 . Then $E(S^2) = \mathbf{s}^2$; that is,

$$S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1} \text{ is an unbiased estimator of } \mathbf{s}^2.$$

Proof: $\frac{(n-1)S^2}{\mathbf{s}^2} \sim \mathbf{c}_{n-1}^2$ by Theorem 2 above.

We also know that for $W \sim \mathbf{c}_{n-1}^2, E(W) = n-1$ so it follows that

$$E\left(\frac{(n-1)S^2}{\mathbf{s}^2}\right) = \frac{(n-1)}{\mathbf{s}^2} E(S^2) = n-1$$

$$\therefore E(S^2) = (n-1) \cdot \frac{\mathbf{s}^2}{n-1} = \mathbf{s}^2$$

Note that it is *not* true that S is an unbiased estimator of \mathbf{s} ; that is $E(S) \neq \mathbf{s}$.

In proving that $E(S^2) = \mathbf{s}^2$ above, we assumed that Y_1, Y_2, \dots, Y_n are independent, identically distributed *normal* random variables. It can be shown that $E(S^2) = \mathbf{s}^2$ whenever Y_1, Y_2, \dots, Y_n are independent, identically (but not necessarily normally) distributed random variables, each with variance \mathbf{s}^2 . We will now derive this more general result.

To prove that $E(S^2) = \mathbf{s}^2$, we need to use the identity

$$\sum_{i=1}^n (y_i - \bar{y})^2 = \sum_{i=1}^n (y_i - \mathbf{m})^2 - n(\bar{y} - \mathbf{m})^2$$

First we give a proof of identity:

$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n [(y_i - \mathbf{m}) + (\mathbf{m} - \bar{y})]^2 \\
&= \sum_{i=1}^n \{(y_i - \mathbf{m})^2 + 2(y_i - \mathbf{m})(\mathbf{m} - \bar{y}) + (\mathbf{m} - \bar{y})^2\} \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 + \sum_{i=1}^n 2(y_i - \mathbf{m})(\mathbf{m} - \bar{y}) + \sum_{i=1}^n (\mathbf{m} - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 + 2(\mathbf{m} - \bar{y}) \sum_{i=1}^n (y_i - \mathbf{m}) + n(\mathbf{m} - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 + 2(\mathbf{m} - \bar{y}) \left\{ \sum_{i=1}^n y_i - \sum_{i=1}^n \mathbf{m} \right\} + n(\mathbf{m} - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 + 2(\mathbf{m} - \bar{y}) \{n\bar{y} - n\mathbf{m}\} + n(\mathbf{m} - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 + 2n(\mathbf{m} - \bar{y})(\bar{y} - \mathbf{m}) + n(\mathbf{m} - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 - 2n(\bar{y} - \mathbf{m})^2 + n(\mathbf{m} - \bar{y})^2
\end{aligned}$$

Thus,
$$\begin{aligned}
\sum_{i=1}^n (y_i - \bar{y})^2 &= \sum_{i=1}^n (y_i - \mathbf{m})^2 - 2n(\bar{y} - \mathbf{m})^2 + n(\mathbf{m} - \bar{y})^2 \\
&= \sum_{i=1}^n (y_i - \mathbf{m})^2 - n(\bar{y} - \mathbf{m})^2
\end{aligned}$$

And now we will prove the theorem that supports dividing by $n-1$.

Theorem: Let Y_1, Y_2, \dots, Y_n be identically and independently distributed random variables with mean \mathbf{m} and finite variance \mathbf{s}^2 .

Define the random variable,
$$S^2 = \frac{\sum_{i=1}^n (Y_i - \bar{Y})^2}{n-1}.$$

Then $E(S^2) = \mathbf{s}^2$.

Proof:

To prove this we will take expected values and use the identity derived above.

$$E\left(\sum_{i=1}^n (Y_i - \bar{Y})^2\right) = E\left\{\sum_{i=1}^n (Y_i - \mathbf{m})^2 - n(\bar{Y} - \mathbf{m})^2\right\}$$

$$= E\left\{\sum_{i=1}^n (Y_i - \mathbf{m})^2\right\} - E\left\{n(\bar{Y} - \mathbf{m})^2\right\}$$

Now,

$$E\left\{\sum_{i=1}^n (Y_i - \bar{Y})^2\right\} = E\{(n-1)S^2\} = (n-1)E(S^2)$$

$$E\left\{\sum_{i=1}^n (Y_i - \mathbf{m})^2\right\} = nV(Y) = n\mathbf{S}^2$$

$$E[n(\bar{Y} - \mathbf{m})^2] = nE[(\bar{Y} - \mathbf{m})^2] = nV(\bar{Y}) = n\left(\frac{\mathbf{S}^2}{n}\right) = \mathbf{S}^2$$

And thus,

$$(n-1)E(S^2) = n\mathbf{S}^2 - \mathbf{S}^2 = (n-1)\mathbf{S}^2,$$

and

$$E(S^2) = \mathbf{S}^2$$

F distribution

The general definition of a random variable with an F distribution was developed by Snedecor around 1920 and named in honor of Fisher. Let $W_1 \sim \mathbf{C}_{n_1}^2$ be independent of $W_2 \sim \mathbf{C}_{n_2}^2$. Then $F = \frac{W_1/n_1}{W_2/n_2}$ is said to have an F distribution with n_1 numerator degrees of freedom and n_2 denominator degrees of freedom. This is often written as F_{n_1, n_2} .