2.4: The Student's t-Distribution and the F Distribution

Textbook: 7.2

Objectives

- Define the Student's t-distribution
- Define the F distribution
- Know when and how to these distributions in appropriate probability calculation scenarios

Motivation

So far in Unit 2, we've learned some properties of three commonly-used statistics that can be used to help summarize or succinctly describe a feature of a random variable.

Value	Useful For…	Parameter	Statistic
Mean	Describing the "typical" value of a quantitative variable	μ	$\bar{x}$
Variance	Describing the variation or "spread" of a quantitative variable about its mean	σ^2	s^2
Proportion	Describing the amount of observations with a certain quality for a qualitative variable	p	$\hat{p}$

We discussed sampling distributions related to these three quantities as well, which allowed us to compute probabilities related to these quantities.

- For a sample of n i.i.d. observations X_i such that $X_i \sim normal(\mu, \sigma)$, then $\bar{x} \sim normal\left(\mu, \frac{\sigma}{\sqrt{n}}\right)$.
- For a sample of $n \geq 25$ i.i.d. observations X_i such that $X_i \sim Bernoulli(p)$, and if $\hat{p} = \frac{\sum X_i}{n}$, then $\frac{\hat{p}-p}{\sqrt{\frac{p(1-p)}{n}}} \sim approximately\ normal(0,1)$.
- For a sample of n i.i.d. observations X_i, then $\frac{(n-1)s^2}{\sigma^2} \sim \chi^2_{df=n-1}$.

But there are certain situations where we might want (or need) more than what we've discovered so far.

Examples

- You wish to determine how long people spend waiting in Tim Horton's drive-thrus across the country, but do not have any information about the population variance.
- A course has two sections, each taught by a different instructor. Suppose the sections use the same final exam. Is there a way to "statistically" compare the variation of final exam scores for section 1 versus section 2?

© 2019-2020 Claudia Mahler

In this last set of notes in Unit 2, we will introduce to relevant distributions that can help us deal with situations similar to the examples presented above: the Student's t-distribution and the F distribution.

The Student's t-Distribution

Definition

Let $Z \sim normal(0,1)$ and let $W \sim \chi^2_{df=\upsilon}$. If Z and W are independent random variables, then the quantity

$$\frac{Z}{\sqrt{\frac{W}{\upsilon}}}$$

is said to follow a <u>Student's t-distribution</u> (<u>t-distribution</u>) with υ degrees of freedom. Like the normal distribution, the family of t-distributions are symmetric and bell-shaped. Unlike the normal distribution, the specific shape of the t-distribution depends on the sample size n. Smaller sample sizes result in "shorter" t-distributions with fatter tails; as $n \to \infty$, the shape of the t-distribution goes to the shape of the normal distribution.

If we let $T = \frac{Z}{\sqrt{\frac{W}{\upsilon}}}$, the PDF of the t-distribution is given as

$$f_T(t) = \frac{\Gamma\left(\frac{(\upsilon+1)}{2}\right)}{\sqrt{\pi\upsilon}\,\Gamma\left(\frac{\upsilon}{2}\right)}\left(1+\frac{t^2}{\upsilon}\right)^{\frac{-(\upsilon+1)}{2}}$$

for $-\infty < t < \infty$.

In addition,

$$E[T] = 0$$

$$VAR[T] = \frac{\upsilon}{\upsilon - 2}$$

Uses

The Student's t-distribution is used in the following situations:

- When you need to compute a probability for the sample mean $\bar{x}$ but only have information about the sample variance s^2 (no info about σ^2)
- When you wish to construct confidence intervals or carry out a hypothesis test for μ when σ^2 is not known
- When you wish to construct confidence intervals or carry out a hypothesis test for the slope and y-intercept parameters in simple linear regression

© 2019-2020 Claudia Mahler

For now, we will focus on its use when dealing with probability calculations for $\bar{x}$.

Consider an i.i.d. sample of $X_i \sim normal(\mu, \sigma^2)$ and the resulting sample mean $\bar{x}$ and sample variance s^2. Consider

$$\frac{\bar{x}-\mu}{s/\sqrt{n}}$$

What is the distribution of $\frac{\bar{x}-\mu}{s/\sqrt{n}}$?

Result:

Note that we usually denote this calculation as

$$T = \frac{\bar{x}-\mu}{s/\sqrt{n}}$$

with $df = n - 1$.

We can now use this result to calculate probabilities for $\bar{x}$ when we do not know σ^2 (and thus have to estimate it with s^2).

© 2019-2020 Claudia Mahler

Example 2.4.1
Tim Hortons puts out a statement claiming that waiting times in their drive-thrus will vary from customer to customer but will do so in accordance with a normal distribution with a mean of 1.02 minutes. You take a random sample of $n = 22$ customers and find $s^2 = 1.56$ minutes. In another sample of the same size, what is the probability that the average waiting time will be between 0.88 minutes and 1.34 minutes?

R function: `pt(x,df)`

```
pt(        , df =    ) - pt(          , df =    )
```

Example 2.4.2
A mathematics textbook publisher states that the prices of their textbooks will vary from textbook to textbook in accordance with a normal distribution with a mean of $250. A student takes a random sample of $n = 15$ textbooks and finds a sample variance of $s^2 = \$50$. In another sample of $n = 15$ textbooks, what is the probability that the average cost of the textbooks is at least $245?

R function: `pt(x,df)`

```
1 - pt(         , df =    )
```

© 2019-2020 Claudia Mahler

The F Distribution

Definition

Let X and Y be independent χ^2-distributed random variables with υ and δ degrees of freedom, respectively. Then

$$\frac{X/\upsilon}{Y/\delta}$$

is said to follow an F distribution with υ degrees of freedom in the numerator and δ degrees of freedom in the denominator. The family of F distributions are a ratio of two χ^2-distributed random variables, each divided by its respective degrees of freedom. The specific shape of an F distribution depends on the two degrees of freedom, but in general, the distributions are right-skewed and $F \geq 0$.

If we let $F = \frac{X/\upsilon}{Y/\delta}$, the PDF of the F distribution is given as

$$f_F(f) = \frac{\Gamma\left(\frac{(\upsilon+\delta)}{2}\right)}{\Gamma\left(\frac{\upsilon}{2}\right)\Gamma\left(\frac{\delta}{2}\right)}\left(\frac{\upsilon}{\delta}\right)^{\frac{\upsilon}{2}} f^{\left(\frac{\upsilon}{2}-1\right)}\left(1+\frac{\upsilon f}{\delta}\right)^{\frac{-(\upsilon+\delta)}{2}}$$

for $f > 0$.

In addition,

$$E[F] = \frac{\delta}{\delta - 2}$$

$$VAR[F] = \frac{2\delta^2(\delta+\upsilon-2)}{\upsilon(\delta-2)^2(\delta-4)}$$

Uses

The F-distribution is used in the following situations:

- When you wish to assess the ratio of two variances to determine if one variance is "significantly" larger than the other
- In situations involving analysis of variance (ANOVA)
- As a test of "global fit" in regression

For our purposes right now, we will be concerned with using the F distribution in situations where we wish to examine a ratio of two variances. Thus, the following is an important result.

© 2019-2020 Claudia Mahler

Consider the ratio

$$\frac{s_1^2\sigma_2^2}{s_2^2\sigma_1^2}$$

Relationship to the Student's t-Distribution
Let

$$T = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

What is the distribution of T^2?

we showed that $T = \frac{\bar{x} - \mu}{s/\sqrt{n}} \sim T_{n-1}$.

$T = \frac{Z}{\sqrt{W/n-1}}$ where $Z \sim$ normal $(0,1)$ and $W \sim \chi^2_{n-1}$

So $T^2 = \left(\frac{Z}{\sqrt{W/n-1}}\right)^2 = \frac{Z^2}{W/n-1}$ } we know that $Z^2 \sim \chi^2_1$

So $T^2 = \frac{Z^2/\underline{1}}{W/\underline{n-1}}$ } ratio of χ^2 over its df } ratio of W over df $n-1$

Result:

© 2019-2020 Claudia Mahler

*Example 2.4.3**
s_1^2 and s_2^2 are sample variances observed from two independent samples of sizes $n_1 = 25$ and $n_2 = 20$, taken from two different populations (each normally distributed). If $\sigma_1^2 = 3\sigma_2^2$, find the value k such that

$$P\left(\frac{s_1^2}{s_2^2} < k\right) = 0.95$$

R function: `qf(p,df1,df2)`

`qf(        , df1 =     , df2 =     )`

Why would we care about the ratio of variances?

If we found out that $\frac{s_1^2}{s_2^2} = 1$, this means that $s_1^2 = s_2^2$, suggesting that the variances of the respective populations are (statistically) the same!

A ratio greater than one suggests that the variance in the first population is greater than the variance in the second population; a ratio less than one suggests that the variance in the first population is less than the variance in the second population. The further this ratio deviates from one in either direction, the stronger evidence we have to suggest that the population variances are not similar!

*Note: all examples marked with an asterisk * are © 2017 Scott Robison*

© 2019-2020 Claudia Mahler