Lecture 28

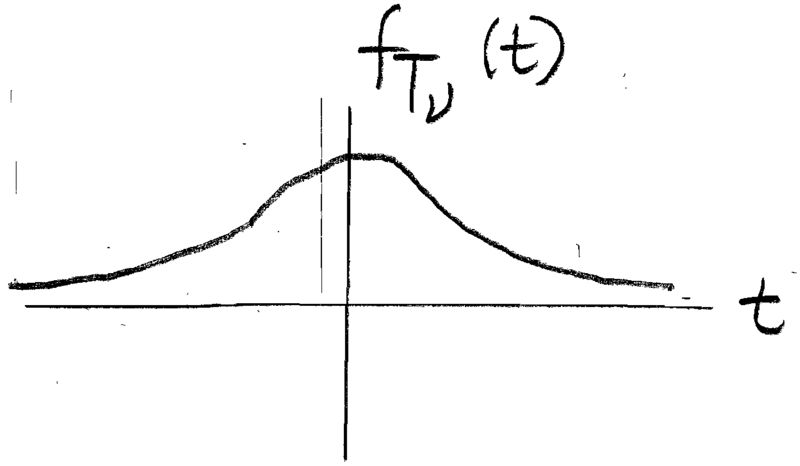
The t-distribution(s) and the Gosset / Student Theorem

The following distribution and its importance was discovered by William Gosset who published under the pseudonym "Student".

Definition

A continuous random variable T_ν is said to have t-distribution with ν degrees of freedom, abbreviated $T_\nu \sim t_\nu$ if its pdf f_{T_ν} is given by

$$f_{T_\nu}(t) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\nu\pi}\;\Gamma(\nu/2)} \frac{1}{\left(1+\frac{t^2}{\nu}\right)^{\frac{\nu+1}{2}}}, \quad -\infty < t < \infty$$



The graph of $f_{T_\nu}(t)$ is like the graph of the standard normal density

$f_Z(z) = \frac{1}{\sqrt{2\pi}} e^{-z^2/2}$ except it doesn't go to zero as fast at ∞ and $-\infty$.

In fact

$$\lim_{\nu \to \infty} f_{T_\nu}(t) = \frac{1}{\sqrt{2\pi}} e^{-t^2/2} \qquad (*)$$

This is why you can get z_α from the last line ($\nu = \infty$) on the back flap of the text.

Note that $f_{T_\nu}(t) = f_{T_\nu}(-t)$

So, if $F_{T_\nu}(t)$ is the cdf of T_ν we have the functional equation

$$F_{T_\nu}(-t) = 1 - F_{T_\nu}(t)$$

and

$$F_{T_\nu}(a) - F_{T_\nu}(-a) = 2F_{T_\nu}(a) - 1$$

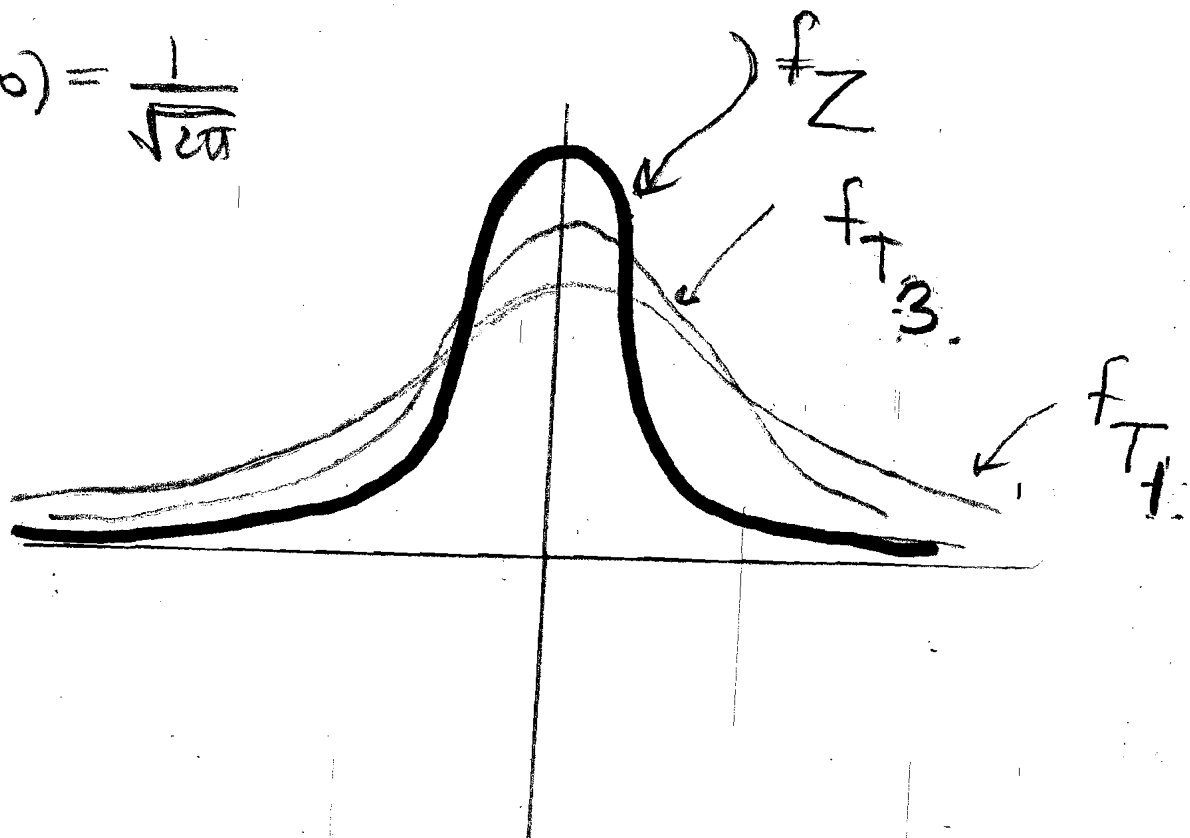
Or in other words, "the handy formula"

$$P(-a \le T_\nu \le a) = 2\,F_{T_\nu}(a) - 1$$

holds just as for the standard normal distribution Z.

Comparison between the different t-curves and the Z-curve

$$f_Z(0) = \frac{1}{\sqrt{2\pi}}$$



$$f_{T_1}(0) = \frac{\Gamma(1)}{\sqrt{\pi}\,\Gamma(\frac{1}{2})} = \frac{1}{\sqrt{\pi}\sqrt{\pi}} = \frac{1}{\pi}$$

$$f_{T_3}(0) = \frac{\Gamma(2)}{\sqrt{2\pi}\,\Gamma(3/2)} = \frac{1}{\sqrt{2\pi}\,\frac{\sqrt{\pi}}{2}} = \frac{2}{\sqrt{2}\,\pi} = \frac{\sqrt{2}}{\pi}$$

$$f_{T_5}(0) = \frac{\Gamma(3)}{\sqrt{3\pi}\,\Gamma(\frac{5}{2})} = \frac{2}{\sqrt{3\pi}\,(\frac{3}{2})(\frac{1}{2})\sqrt{\pi}} = \frac{8}{3\sqrt{3}\,\pi}$$

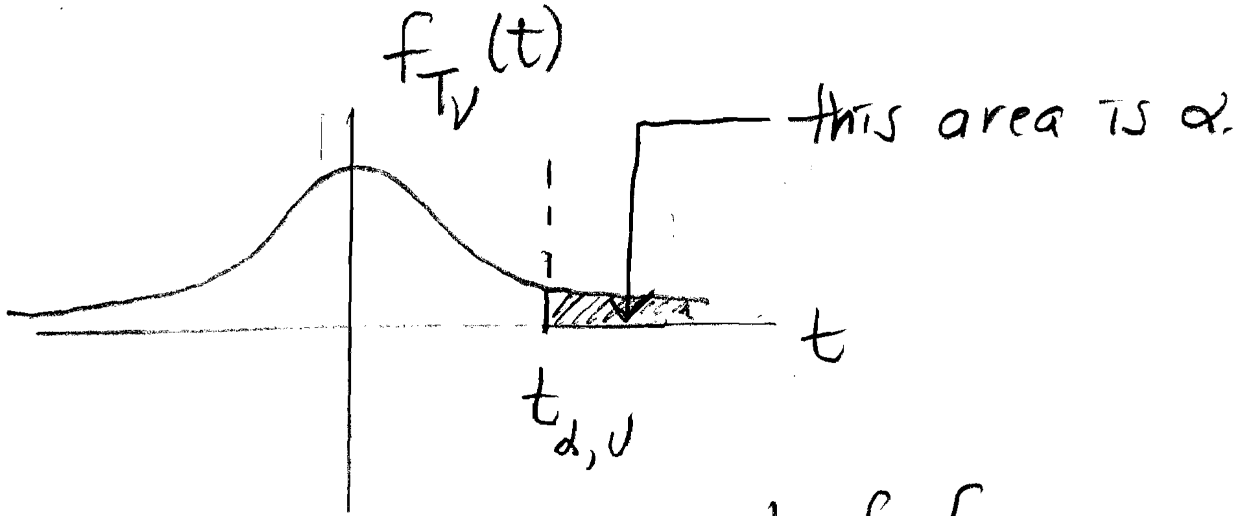
The Critical Values of T_ν

Definition

Let α be a real number between 0 and 1. Then the α-th critical value $t_{\alpha,\nu}$ for T_ν is the number satisfying

$$P(T_\nu \geq t_{\alpha,\nu}) = \alpha$$

Geometrically we have



So the area under the graph of f_{T_ν} to the right of the vertical line $t = t_{\alpha,\nu}$ is α.

The critical values $t_{\alpha;\nu}$ are on the back flap of the text

ble A.5 Critical Values for *t* Distributions

ν \ α	.10	.05	.025	.01
1	3.078	6.314	12.706	31.821
2	1.886	2.920	4.303	6.965
3	1.638	2.353	3.182	4.541
4	1.533	2.132	2.776	3.747
5	1.476	2.015	2.571	[illegible]
6	1.440	1.943	2.447	[illegible]
7	1.415	1.895	2.365	[illegible]
8	1.397	1.860	2.306	[illegible]
9	1.383	1.833	2.262	[illegible]
10	1.372	1.812	2.228	[illegible]
11	1.363	1.796	2.201	[illegible]
12	1.356	1.782	2.179	[illegible]
13	1.350	1.771	2.160	[illegible]
14	1.345	1.761	2.145	[illegible]
15	1.341	1.753	2.131	[illegible]
16	1.337	1.746	2.120	[illegible]
17	1.333	1.740	2.110	[illegible]
18	1.330	1.734	2.101	[illegible]
19	1.328	1.729	2.093	[illegible]

$$t_{9,.05} = 1.833$$

Why is the t-distribution important?

Gosset / Student's great observation was (probably Fisher proved this - see the Wikipedia article)

<u>Theorem FG</u>

Suppose $Z \sim N(0,1)$ and $V \sim \chi^2(m)$ and Z and V are independent. Put

$$T = \frac{Z}{\sqrt{V/m}}.$$

Then $T \sim t_m$

<u>Remark</u> Of course the main point was to realize that one should look at the above ratio. In fact it seems Gosset left out the $\sqrt{m}$ - from Wikipedia.

Now we do we want to look at the above ratio? What was Gosset's idea?

Well, in the formula for the Z- confidence interval we were led to

$$Z = \frac{\overline{X} - \mu}{\sigma/\sqrt{n}}.$$

But what if <u>we don't know</u> σ.

<u>Idea</u>

Replace σ by its point estimator S so we replace

$$\frac{\overline{X} - \mu}{\sigma/\sqrt{n}} \quad \text{by} \quad \frac{\overline{X} - \mu}{S/\sqrt{n}}$$

From Theorem FG (which is moderately hard to prove) it is an exercise in fractions to prove

Theorem A

Suppose $X_1, X_2, \ldots, X_n$ is a random sample from a normal population with mean μ and variance σ^2. Then

$$T = \frac{\overline{X} - \mu}{S/\sqrt{n}} \sim t_{n-1} \qquad (**)$$

We will need the results from page 5 of Lecture 25

(i) $\overline{X} \sim N(\mu, \frac{\sigma^2}{n})$

(ii) $\frac{n-1}{\sigma^2} S^2 \sim \chi^2(n-1)$

(iii) $\overline{X}$ and S^2 are independent

Now we will prove (**).

The idea is we want to change $\bar{X}-\mu$ into Z so we have to divide the numerator by $\sigma/\sqrt{n}$ – so we also have to divide the denominator by $\sigma/\sqrt{n}$ so that the fraction has the same value

$$\frac{\bar{X}-\mu}{(S/\sqrt{n})} = \frac{(\bar{X}-\mu)/(\sigma/\sqrt{n})}{(S/\sqrt{n})/(\sigma/\sqrt{n})} = \frac{Z}{\left(\frac{\frac{S}{\cancel{\sqrt{n}}}}{\frac{\sigma}{\cancel{\sqrt{n}}}}\right)}$$

$$= \frac{Z}{\left(\frac{S}{\sigma}\right)} = \frac{Z}{\sqrt{\frac{S^2}{\sigma^2}}}$$

$$= \frac{Z}{\sqrt{\frac{n-1}{\sigma^2} S^2 \frac{1}{n-1}}}$$

Put $V = \frac{n-1}{\sigma^2} S^2$ so $V \sim \chi^2(n-1)$.

Hence we obtain

$$\frac{\bar{X} - \mu}{\left(S/\sqrt{n}\right)} = \frac{Z}{\sqrt{V/n-1}}$$

But by this is exact the right ratio to get a t-distribution with n-1 degrees of freedom

$\square$

Remark

Before Gosset statisticians assumed that $\frac{\bar{X} - \mu}{S/\sqrt{n}}$ was standard normal.

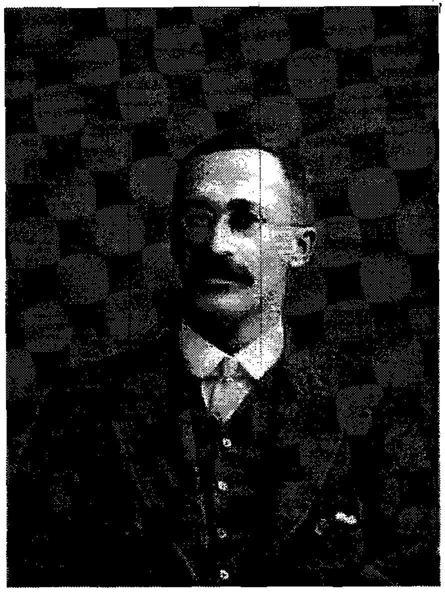
This is approximateley true if n is large but far from true if n is not large.

12

William Sealy Gosset

From Wikipedia, the free encyclopedia

William Sealy Gosset



Student in 1908

Born	June 13, 1876 Canterbury, Kent, England
Died	October 16, 1937 (aged 61) Beaconsfield, Buckinghamshire, England

William Sealy Gosset (June 13, 1876–October 16, 1937) is famous as a statistician, best known by his pen name ***Student*** and for his work on Student's t-distribution.

Born in Canterbury, England to Agnes Sealy Vidal and Colonel Frederic Gosset, Gosset attended Winchester College before reading chemistry and mathematics at New College, Oxford. On graduating in 1899, he joined the Dublin brewery of Arthur Guinness & Son.

Guinness was a progressive agro-chemical business and Gosset would apply his statistical knowledge both in the brewery and on the farm—to the selection of the best yielding varieties of barley. Gosset acquired that knowledge by study, trial and error and by spending two terms in 1906–7 in the biometric laboratory of Karl Pearson. Gosset and Pearson had a good relationship and Pearson helped Gosset with the mathematics of his papers. Pearson helped with the 1908 papers but he had little appreciation of their importance. The papers addressed the brewer's concern with small samples, while the biometrician typically had hundreds of observations and saw no urgency in developing small-sample methods.

Another researcher at Guinness had previously published a paper containing trade secrets of the Guinness brewery. To prevent further disclosure of confidential information, Guinness prohibited its employees from publishing any papers regardless of the contained information. This meant that Gosset was unable to publish his works under his own name. He therefore used the pseudonym *Student* for his publications to avoid their detection by his employer. Thus his most famous achievement is now referred to as Student's t-distribution, which might otherwise have been Gosset's t-distribution.

Gosset had almost all of his papers including *The probable error of a mean* published in Pearson's journal *Biometrika* using the pseudonym *Student*. However, it was R. A. Fisher who appreciated the importance of Gosset's small-sample work, after Gosset had written to him to say *I am sending you a copy of Student's Tables as you are the only man that's ever likely to use them!*. Fisher believed that Gosset had effected a "logical revolution". Ironically the *t*-statistic for which Gosset is famous was actually Fisher's creation. Gosset's statistic was $z = t/\sqrt{(n-1)}$. Fisher introduced the *t*-form because it fit in with his theory of degrees of freedom. Fisher was also responsible for the applications of the *t*-distribution to regression.

Although introduced by others, Studentized residuals are named in Student's honor because, like the problem that led to Student's t-distribution, the idea of adjusting for estimated standard deviations is central to that concept.

Gosset's interest in barley cultivation led him to speculate that design of experiments should aim, not only at improving the average yield, but also at breeding varieties whose yield was insensitive (robust) to variation in soil and climate. This principle only occurs in the later thought of Fisher and then in the work of Genichi Taguchi in the 1950s.

In 1935, he left Dublin to take up the position of Head Brewer, in charge of the scientific side of production, at a new Guinness brewery at Park Royal in North West London. He died in Beaconsfield, England of a heart attack.

Gosset was a friend of both Pearson and Fisher, an achievement, for each had a massive ego and a loathing for the other.[citation needed] Gosset was a modest man who cut short an admirer with the comment that "Fisher would have discovered it all anyway."

I should say so!!

Bibliography

- *The application of the law of error to the work of the Brewery* (1904, nota interna presso *Guinness*)
- "On the error of counting with hæmacytometer". *Biometrika* **5** (3): 351–360. February 1907.
- "The probable error of a mean" (http://www.york.ac.uk/depts/maths/histstat/student.pdf) . *Biometrika* **6** (1): 1–25. March 1908. doi:10.1093/biomet/6.1.1 (http://dx.doi.org/10.1093%2Fbiomet%2F6.1.1) . http://www.york.ac.uk/depts/maths/histstat/student.pdf.
- "Probable error of a correlation coefficient". *Biometrika* **6** (2/3): 302–310. September 1908. doi:10.1093/biomet/6.2-3.302 (http://dx.doi.org/10.1093%2Fbiomet%2F6.2-3.302) .
- "The distribution of the means of samples which are not drawn at random". *Biometrika* **7** (1/2): 210–214. July–October 1909. doi:10.1093/biomet/7.1-2.210 (http://dx.doi.org/10.1093%2Fbiomet%2F7.1-2.210) .
- "An experimental determination of the probable error of Dr Spearman's correlation coefficients". *Biometrika* **13** (2/3): 263–282. July 1921. doi:10.1093/biomet/13.2-3.263 (http://dx.doi.org/10.1093%2Fbiomet%2F13.2-3.263) .
- "Review of Statistical Methods for Research Workers (R. A. Fisher)" (http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/student.htm) . *Eugenics Review* **18**: 148–150. 1926. http://www.economics.soton.ac.uk/staff/aldrich/fisherguide/student.htm.
- "On Student's 1908 Article "The Probable Error of a Mean"(S.L.Zabell)" (http://cda.mrs.umn.edu/~jongmink/Stat2611/s1.pdf) . *Journal of the American Statistical Association* **Vol. 103, No. 481**: 1–7. March 2008. http://cda.mrs.umn.edu/~jongmink/Stat2611/s1.pdf.
- *'Student's' Collected Papers* (edited by E.S. Pearson and John Wishart, with a foreword by Launce McMullen), London: Biometrika Office. (1942)

Biographies

- E. S. Pearson (1990) *'Student', A Statistical Biography of William Sealy Gosset,* Edited and Augmented by R. L. Plackett with the Assistance of G. A. Barnard, Oxford: University Press.
- E. S. Pearson, "'Student' as Statistician, *Biometrika* Vol. 30, No. 3/4 (Jan., 1939), pp. 210-250.

External links

- Biography by Heinz Kohler (http://www.swlearning.com/quant/kohler/stat/biographical_sketches/bio12.1.html)
- Tales of Statisticians by E. Bruce Brooks (http://www.umass.edu/wsp/statistics/tales/gosset.html)
- Student's T Distribution (http://www-stat.stanford.edu/~naras/jsm/TDensity/TDensity.html)
- Earliest known uses of some of the words of mathematics: S (http://jeff560.tripod.com/s.html) under the heading of "Student's *t*-distribution", describes briefly how Student's *z* became *t*.
- O'Connor, John J.; Robertson, Edmund F., "William Sealy Gosset" (http://www-history.mcs.st-andrews.ac.uk/Biographies/Gosset.html) , *MacTutor History of Mathematics archive*, University of St Andrews, http://www-history.mcs.st-andrews.ac.uk/Biographies/Gosset.html.

Retrieved from "http://en.wikipedia.org/wiki/William_Sealy_Gosset"
Categories: 1876 births | 1937 deaths | People from Canterbury | Old Wykehamists | Alumni of New College, Oxford | British mathematicians | Deaths from myocardial infarction | English statisticians