STA257: Probability and Statistics 1

Instructor: Katherine Daignault

Department of Statistical Sciences
University of Toronto

Week 12

Outline

Outline

Distributions for future courses

- Our last week of material will focus on introducing some new distributions of random variables that will feature prominently in future statistics courses.
- We can derive them with techniques encountered in this course
- We will also talk about how these new distributions are used to represent the distribution of sample means and variances.
- These distributions are most commonly used to perform hypothesis tests regarding the value of a parameter or the relationship between variables.
  - so like what our gambler was doing to determine whether his die was fair.

Chi-Square Distribution

- Our first distribution is one that we have already encountered in Week 5
- There, we saw that if we take a square transformation of a standard Normal, we get the Chi Square distribution.
- Recall that if $Z \sim N(0,1)$ and $X = Z^2$ then X has PDF

$$f_X(x) = \frac{x^{-1/2}}{\sqrt{2\pi}} e^{-x/2}, \quad x \geq 0$$

  which is equivalent to $X \sim Gamma(0.5, 0.5)$.
- Here we say that $X \sim \chi^2_1$, where 1 is the parameter that dictates the shape of the distribution.
  - The parameter here is called the degrees of freedom

Chi-Square Distribution

- It turns out we can generalize the chi square random variable, by considering the sum of independent χ_1^2 random variables.

General Chi Square

If $U_1, U_2, \ldots, U_n$ are independent chi-square random variables with 1 degree of freedom, the distribution of $V = U_1 + U_2 + \cdots + U_n$ is called the chi-square distribution with n degrees of freedom, and is denoted by χ_n^2.

Proof:

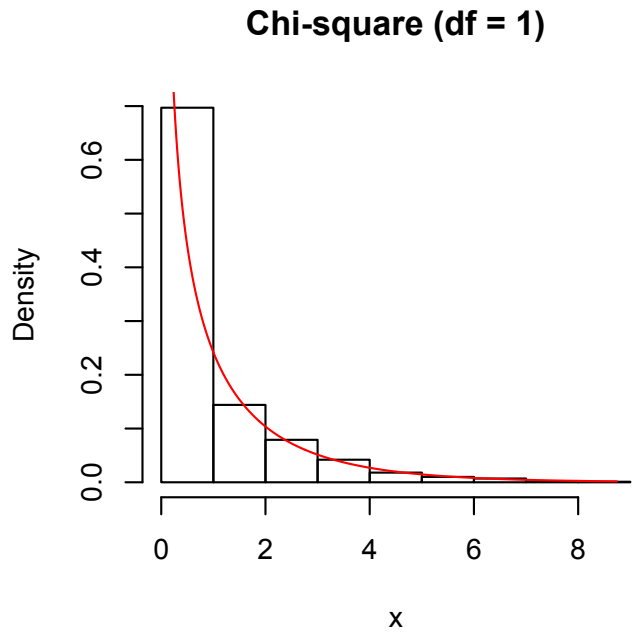
Chi-Square Distribution

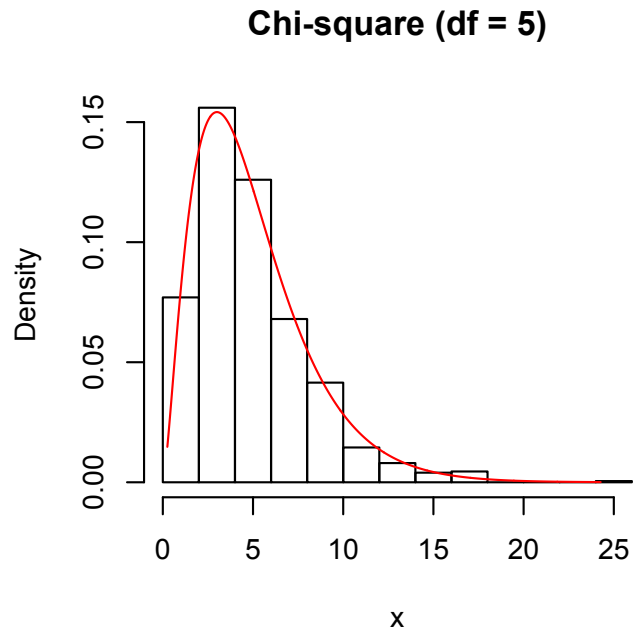
- So we see now that the degrees of freedom represents how many independent χ^2_1 we are summing together.
- A natural consequence of this is that we can easily find the distribution for the sum of any independent chi-square random variables.
  - if $U \sim \chi^2_n$ and $V \sim \chi^2_m$ then $U + V \sim \chi^2_{m+n}$.
  - this can be proved in the same way as the previous result.
- Because the Gamma and Chi-square distributions are so related to each other, we can use this to find the PDF of χ^2_n random variable

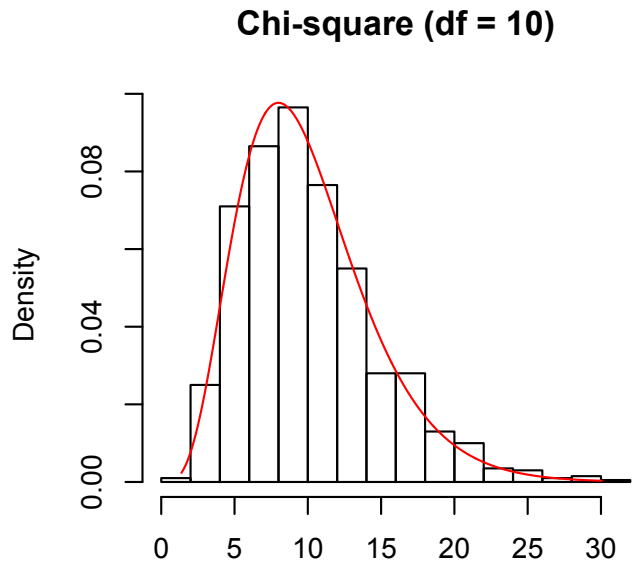
Exercise - Give it a try!

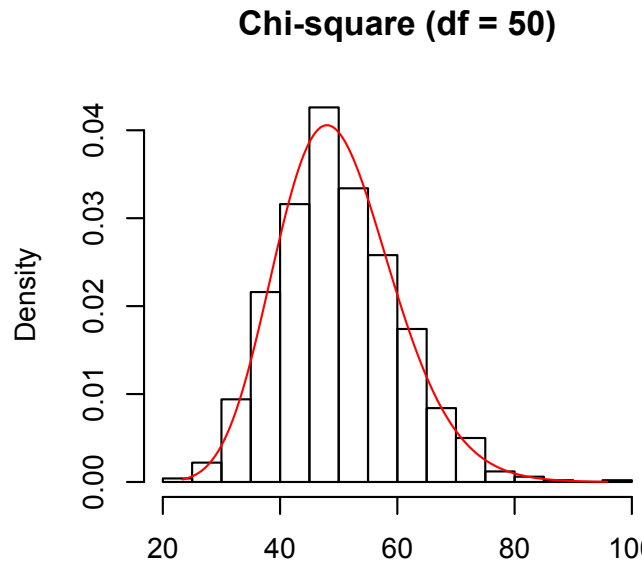
Find the PDF of a χ^2_n random variable.

Visualizing the Chi-square distribution









Chi-Square distribution - Uses

- Since a chi-square random variable only takes on positive values, and is built from squaring a standard Normal, it is sometimes used to model the behaviour of sample variances (which we will get into later)
- More often, it is used to perform a variety of hypothesis tests, such as:
  - Chi-square test for independence in contingency tables
  - Chi-square test for goodness of fit of observed data to a theoretical model
  - Log-rank test for comparing survival rates between two groups.
- In these applications, we are often trying to find the probability of being more extreme than a certain value of the chi-square.

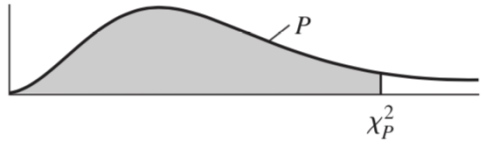
Example: Variation in Production

Suppose there is variation in the weight of dice being produced in a factory. The company wants to model this variability using a chi-squared distribution with 3 degrees of freedom. As a quality control measure, the company will destroy all dice whose variation is in upper 10% of the distribution of weight variations. What percentile cutoff on variation should be used?

- Our random variable, representing the variation in dice weight, is $X \sim \chi^2_3$.
- We want to find the percentile representing the upper 10% of the distribution, or equivalently the lower 90% of the distribution.
- This corresponds to finding the value c such that $P(X \leq c) = 0.9$
- How do we find this probability?

Chi-square Table of Percentiles

TABLE 3 Percentiles of the χ^2 Distribution—Values of χ^2_P Corresponding to P



df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19
11	2.60	3.05	3.82	4.57	5.58	17.28	19.68	21.92	24.73	26.76
12	3.07	3.57	4.40	5.23	6.30	18.55	21.03	23.34	26.22	28.30
13	3.57	4.11	5.01	5.89	7.04	19.81	22.36	24.74	27.69	29.82
14	4.07	4.66	5.63	6.57	7.79	21.06	23.68	26.12	29.14	31.32
15	4.60	5.23	6.26	7.26	8.55	22.31	25.00	27.49	30.58	32.80
16	5.14	5.81	6.91	7.96	9.31	23.54	26.30	28.85	32.00	34.27
18	6.26	7.01	8.23	9.39	10.86	25.99	28.87	31.53	34.81	37.16
20	7.43	8.26	9.59	10.85	12.44	28.41	31.41	34.17	37.57	40.00
24	9.89	10.86	12.40	13.85	15.66	33.20	36.42	39.36	42.98	45.56
30	13.79	14.95	16.79	18.49	20.60	40.26	43.77	46.98	50.89	53.67
40	20.71	22.16	24.43	26.51	29.05	51.81	55.76	59.34	63.69	66.77
60	35.53	37.48	40.48	43.19	46.46	74.40	79.08	83.30	88.38	91.95
120	83.85	86.92	91.58	95.70	100.62	140.23	146.57	152.21	158.95	163.64

T distribution

- The chi-square is also used to derive other helpful distributions.
- Our next distribution, if you were to plot the density, looks a lot like a Normal distribution (i.e. symmetric about the mean, bell-shaped).
- The only difference between the Normal and the T distribution is that the T will have heavier tails than the Normal.
  - heavier tails means that there is a higher chance with the T distribution to observe extreme values (far from the mean) than compared to the Normal.

T distribution

If $Z \sim N(0,1)$ and $U \sim \chi^2_n$, and U and Z are independent, then the distribution of $Z/\sqrt{U/n}$ is called a T distribution with n degrees of freedom.

T distribution

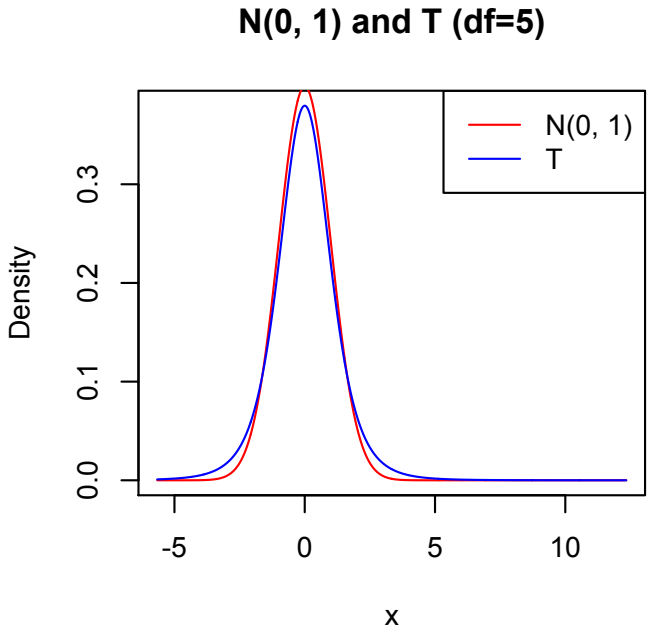
- The heavier tails arise due to the chi-square distribution in the denominator of the product, as it injects extra variation into the Normal distribution.
- Because of this, the shape of the density function is determined by the degrees of freedom of the chi-square random variable.
- The PDF of the T distribution can be found (and is assigned as homework) to be:

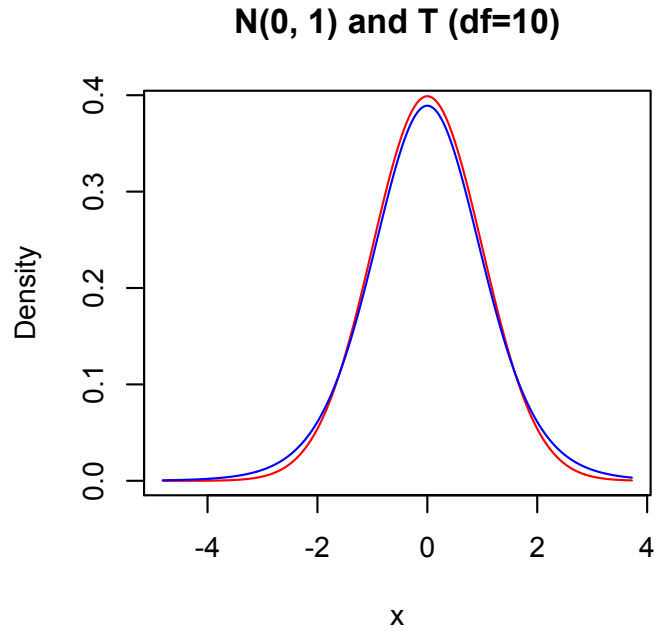
PDF of T distribution

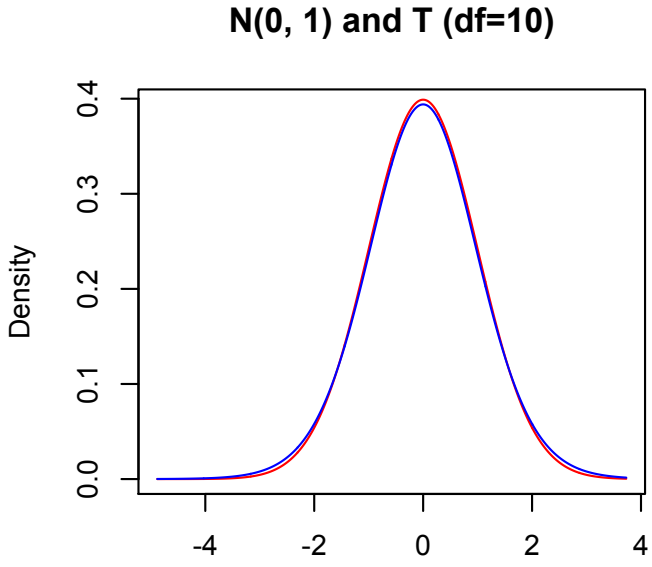
The density function of the T distribution with n degrees of freedom is

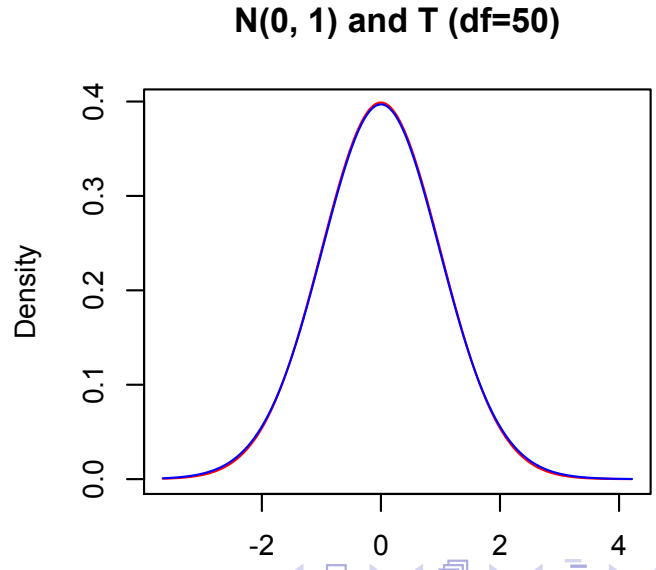
$$f(t) = \frac{\Gamma\left((n+1)/2\right)}{\sqrt{n\pi}\Gamma\left(n/2\right)}\left(1+\frac{t^2}{n}\right)^{-(n+1)/2}$$

Visualizing the T distribution







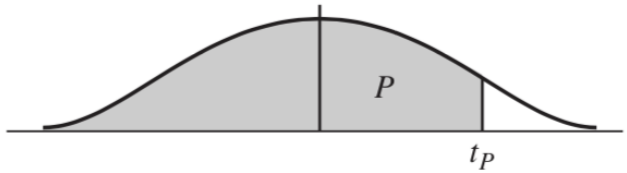


T distribution - Uses

- Since the T distribution is created through a ratio of a $N(0,1)$ and χ^2_n random variable, it is often used to model the behaviour of standardized quantities.
- We will see a bit later how it can be used in this situation.
- It comes up a lot again in situations regarding inference about a parameter (subject of STA261):
  - hypothesis tests about population means when the population variance is unknown
  - confidence intervals and p-values regarding population means when variance is unknown
  - constructing prediction intervals for unknown means
- Since we often need to find probabilities/percentiles from a T distribution, we have another table that can be used, like the chi-square.

T table of Probabilities

TABLE 4 Percentiles of the t Distribution



df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169
11	.260	.540	.876	1.363	1.796	2.201	2.718	3.106
12	.259	.539	.873	1.356	1.782	2.179	2.681	3.055
13	.259	.538	.870	1.350	1.771	2.160	2.650	3.012
14	.258	.537	.868	1.345	1.761	2.145	2.624	2.977
15	.258	.536	.866	1.341	1.753	2.131	2.602	2.947
16	.258	.535	.865	1.337	1.746	2.120	2.583	2.921
17	.257	.534	.863	1.333	1.740	2.110	2.567	2.898
18	.257	.534	.862	1.330	1.734	2.101	2.552	2.878
19	.257	.533	.861	1.328	1.729	2.093	2.539	2.861
20	.257	.533	.860	1.325	1.725	2.086	2.528	2.845

F distribution

- Our last new distribution for this course, which is again very useful in future courses, is called the F distribution.
- In this case, it is built using two chi-square random variables
- But it is also related to the T distribution, and can in some cases be expressed as very specific transformations of a T random variable.

F distribution

Let U and V be independent chi-square random variables with m and n degrees of freedom, respectively. The distribution of

$$W = \frac{U/m}{V/n}$$

is called the F distribution with m and n degrees of freedom and is denoted by $F_{m,n}$

Density of F distribution

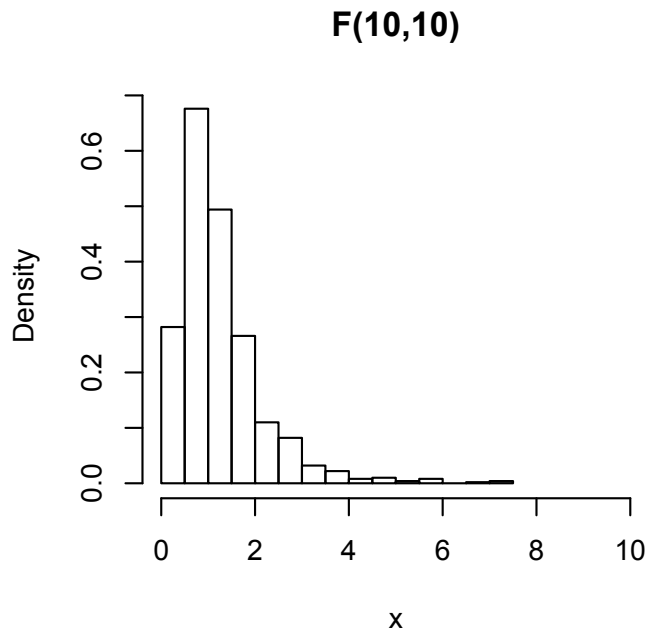
PDF of F distribution

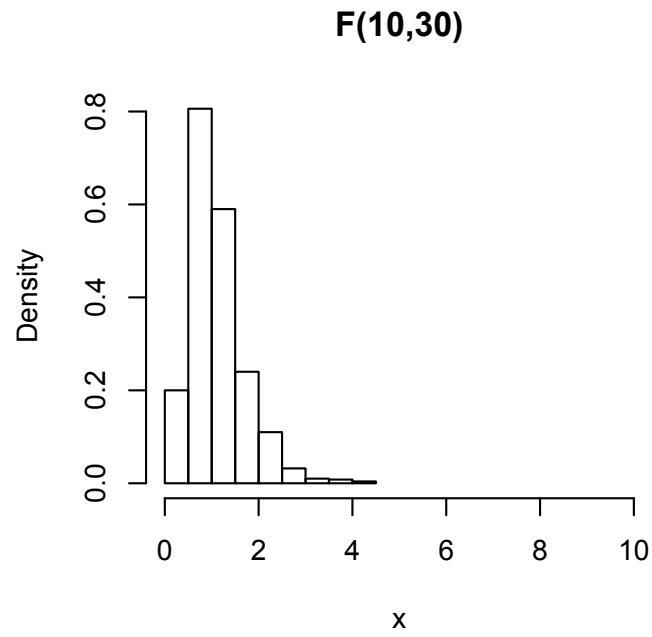
The density function of W is given by

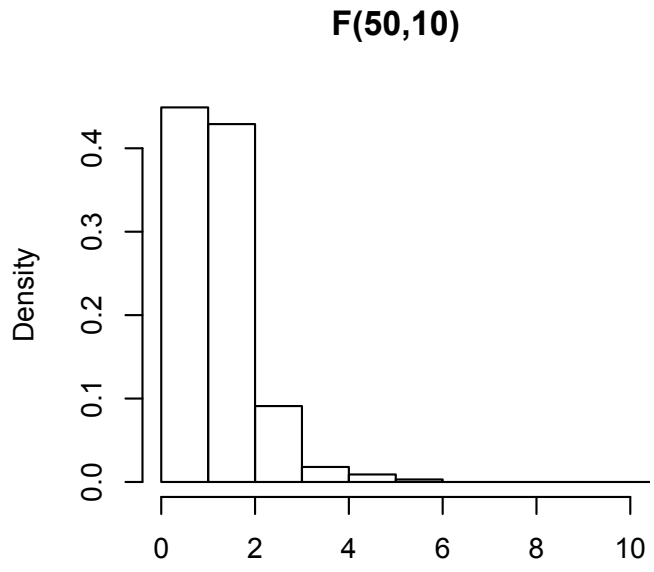
$$f(w) = \frac{\Gamma(m+n/2)}{\Gamma(m/2)\Gamma(n/2)} \left(\frac{m}{n}\right)^{m/2} \left(1+\frac{m}{n}w\right)^{-(m+n)/2}, \quad w \geq 0$$

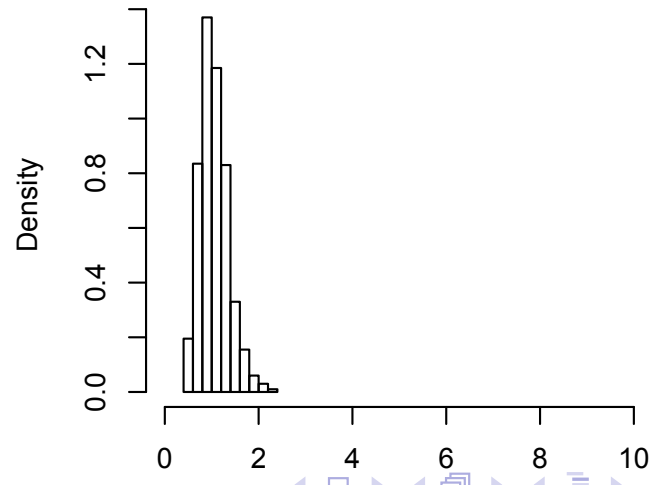
Proof:

Visualizing the F Distribution







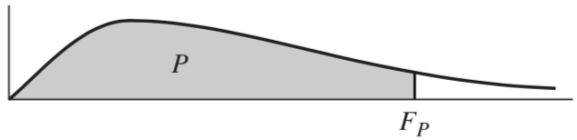


F Distribution - Uses

- Once again, since the F distribution is a ratio of Chi-square random variables, it can be used to model the behaviour of the ratio of variances.
- This is often useful to run hypothesis tests, such as
  - conducting an Analysis of Variance procedure in linear regression to determine how much variation has been explained by your model
  - an F-test to determine whether two population means are equal
- In hypothesis testing procedures, we often need to find the probability of observing data more extreme than our one sample.
- Like the other distributions mentioned today, we can find such probabilities/percentiles with the use of a table.

F table of Percentiles

TABLE 5 Percentiles of the F Distribution: $F_{.90}(n_1, n_2)$



$F_{.90}(n_1, n_2)$

n_1 = degrees of freedom for numerator

n_2 = degrees of freedom for denominator

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	39.86	49.50	53.59	55.83	57.24	58.20	58.91	59.44	59.86	60.19	60.71	61.22	61.74	62.00	62.26	62.53	62.79	63.06	63.33
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39	9.41	9.42	9.44	9.45	9.46	9.47	9.47	9.48	9.49
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23	5.22	5.20	5.18	5.18	5.17	5.16	5.15	5.14	5.13
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92	3.90	3.87	3.84	3.83	3.82	3.80	3.79	3.78	3.76
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30	3.27	3.24	3.21	3.19	3.17	3.16	3.14	3.12	3.10
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94	2.90	2.87	2.84	2.82	2.80	2.78	2.76	2.74	2.72
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70	2.67	2.63	2.59	2.58	2.56	2.54	2.51	2.49	2.47
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.50	2.50	2.46	2.42	2.40	2.38	2.36	2.34	2.32	2.29
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42	2.38	2.34	2.30	2.28	2.25	2.23	2.21	2.18	2.16
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32	2.28	2.24	2.20	2.18	2.16	2.13	2.11	2.08	2.06
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25	2.21	2.17	2.12	2.10	2.08	2.05	2.03	2.00	1.97
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19	2.15	2.10	2.06	2.04	2.01	1.99	1.96	1.93	1.90
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14	2.10	2.05	2.01	1.98	1.96	1.93	1.90	1.88	1.85
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10	2.05	2.01	1.96	1.94	1.91	1.89	1.86	1.83	1.80
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06	2.02	1.97	1.92	1.90	1.87	1.85	1.82	1.79	1.76
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03	1.99	1.94	1.89	1.87	1.84	1.81	1.78	1.75	1.72
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00	1.96	1.91	1.86	1.84	1.81	1.78	1.75	1.72	1.69
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98	1.93	1.89	1.84	1.81	1.78	1.75	1.72	1.69	1.66
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96	1.91	1.86	1.81	1.79	1.76	1.73	1.70	1.67	1.63
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94	1.89	1.84	1.79	1.77	1.74	1.71	1.68	1.64	1.61
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92	1.87	1.83	1.78	1.75	1.72	1.69	1.66	1.62	1.59
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90	1.86	1.81	1.76	1.73	1.70	1.67	1.64	1.60	1.57
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89	1.84	1.80	1.74	1.72	1.69	1.66	1.62	1.59	1.55
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88	1.83	1.78	1.73	1.70	1.67	1.64	1.61	1.57	1.53

F table of Percentiles

TABLE 5 Percentiles of the F Distribution: $F_{.95}(n_1, n_2)$ (Continued)

n_1 = degrees of freedom for numerator

n_2 \ n_1	1	2	3	4	5	6	7	8	9	10	12	15	20	24	30	40	60	120	∞
1	161.4	199.5	215.7	224.6	230.2	234.0	236.8	238.9	240.5	241.9	243.9	245.9	248.0	249.1	250.1	251.1	252.2	253.3	254.3
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.38	19.40	19.41	19.43	19.45	19.45	19.46	19.47	19.48	19.49	19.50
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79	8.74	8.70	8.66	8.64	8.62	8.59	8.57	8.55	8.53
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96	5.91	5.86	5.80	5.77	5.75	5.72	5.69	5.66	5.63
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74	4.68	4.62	4.56	4.53	4.50	4.46	4.43	4.40	4.36
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06	4.00	3.94	3.87	3.84	3.81	3.77	3.74	3.70	3.67
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64	3.57	3.51	3.44	3.41	3.38	3.34	3.30	3.27	3.23
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35	3.28	3.22	3.15	3.12	3.08	3.04	3.01	2.97	2.93
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14	3.07	3.01	2.94	2.90	2.86	2.83	2.79	2.75	2.71
10	4.96	4.10	3.71	3.48	3.83	3.22	3.14	3.07	3.02	2.98	2.91	2.85	2.77	2.74	2.70	2.66	2.62	2.58	2.54
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85	2.79	2.72	2.65	2.61	2.57	2.53	2.49	2.45	2.40
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75	2.69	2.62	2.54	2.51	2.47	2.43	2.38	2.34	2.30
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67	2.60	2.53	2.46	2.42	2.38	2.34	2.30	2.25	2.21
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60	2.53	2.46	2.39	2.35	2.31	2.27	2.22	2.18	2.13
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54	2.48	2.40	2.33	2.29	2.25	2.20	2.16	2.11	2.07
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49	2.42	2.35	2.28	2.24	2.19	2.15	2.11	2.06	2.01
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45	2.38	2.31	2.23	2.19	2.15	2.10	2.06	2.01	1.96
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41	2.34	2.27	2.19	2.15	2.11	2.06	2.02	1.97	1.92
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38	2.31	2.23	2.16	2.11	2.07	2.03	1.98	1.93	1.88
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35	2.28	2.20	2.12	2.08	2.04	1.99	1.95	1.90	1.84
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32	2.25	2.18	2.10	2.05	2.01	1.96	1.92	1.87	1.81
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30	2.23	2.15	2.07	2.03	1.98	1.94	1.89	1.84	1.78
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27	2.20	2.13	2.05	2.01	1.96	1.91	1.86	1.81	1.76
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25	2.18	2.11	2.03	1.98	1.94	1.89	1.84	1.79	1.73

n_2 = degrees of freedom for denominator

Moving Towards Data

- We won't use the F distribution in this course, but it is important to have seen it once before moving on.
- The T distribution and Chi-square, however, will be used in our final topic of the term.
- We have slowly been starting to talk about collecting data, and developing limit results for use when collecting actual data.
- Therefore, we will end the course with some results regarding the sample mean and variance, which will be extremely useful in future courses.

Outline

Random Variables for Sample Quantities

- We have begun to think about the collection of a sample of data from some larger population.
- Last week, we discussed that the more data we collect, the closer we get to approximating certain sample quantities with a Normal distribution.
- Suppose the random variables $X_1, \ldots, X_n$ representing the possible data values we could get are all being drawn independently from a $N(\mu, \sigma^2)$ distribution.
- The sample quantities we will work with today are
  - the sample mean: $\bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i$
  - the sample variance: $S^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})^2$
- We will talk now about the joint and marginal distributions of $\bar{X}$ and S^2.

Working with the Sample Mean

- In Week 11, we saw that a sum of independent RVs with common mean 0 and variance, when scaled by the standard deviation of the sum, will be $N(0,1)$ for sufficiently large n.
  - i.e. $Z_n = \frac{\sum_{i=1}^{n} X_i}{\sigma\sqrt{n}} \sim N(0,1)$
- Previously, we saw that if the X_i did not have mean 0, we could force it to have mean 0 by subtracting the mean from each X_i.
  - i.e. $Z_n = \frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}} \sim N(0,1)$
- By dividing the top and bottom of this by n, we can achieve a CLT result for the sample mean:

$$Z_n = \frac{1/n}{1/n}\frac{\sum_{i=1}^{n}(X_i-\mu)}{\sigma\sqrt{n}} = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

- If we were to back solve for $\bar{X}$, we would find that $\bar{X} \sim N(\mu, \sigma^2/n)$

Example: Bottle Filing Machine

A bottling machine dispenses liquid at an average μ ounces per bottle. It has been observed that amount dispensed is $N(\mu, 1)$. A sample of 9 bottles is randomly selected and the amount of liquid in each bottle is measured. What is the probability that the sample mean will be within 0.3 ounces of the true μ?

- Let $Y_1, \ldots, Y_9$ be the ounces in each bottle sampled, and each $Y_i \sim N(\mu, 1)$
- Therefore, we know that also $\bar{Y} = \frac{1}{9}\sum_{i=1}^{9} Y_i \sim N(\mu, 1/9)$
- When we say "within 0.3 of the true mean", we are referring to $\left|\bar{Y} - \mu\right| \leq 0.3$
- Therefore the probability we are asked to find is

$$P\left(\left|\bar{Y} - \mu\right| \leq 0.3\right) = P(-0.3 \leq \bar{Y} - \mu \leq 0.3)$$

Example: Bottle Filing Machine (cont.)

- From here, we just need to recognize that we almost have a situation where we might use the CLT and the standard Normal table.
- We just need to divide by the standard deviation of the sample mean, $\frac{1}{\sqrt{9}}$
- When we do that,

$$P\left(\left|\bar{Y}-\mu\right| \leq 0.3\right) = P\left(\frac{-0.3}{1/3} \leq Z \leq \frac{0.3}{1/3}\right) = P(-0.9 \leq Z \leq 0.9)$$

- Using symmetry of the Normal distribution, we find

$$P(-0.9 \leq Z \leq 0.9) = 1-2P(Z \leq 0.9) = 1-2(0.8159) = 0.63$$

Exercise - Give it a try!

In the bottle filling example, how many bottles should be included in the sample if we wish $\bar{Y}$ to be within 0.3 ounces of μ with probability 0.95?

Working with the Sample Variance

- We have already easily found that the marginal distribution of the sample mean is $N(\mu, \sigma^2/n)$, so we can now turn to the sample variance.
- Recall that the sample variance takes the form $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$
- Since this features both X_i and $\bar{X}$, we need to understand the relationship between these random variables.
- More specifically, S^2 is essentially a linear combination of random variables $X_i - \bar{X}$, so really we need to understand the relationship between these and $\bar{X}$
- Therefore we need to look into the joint distribution of all $X_i - \bar{X}$ and $\bar{X}$

Joint distribution of all $X_i - \bar{X}$ and $\bar{X}$

Theorem

The random variable $\bar{X}$ and the vector of random variables $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$ are independent.

- The proof of this is given in the book, and relies on some "algebraic trickery to show that the joint MGF factors into the MGF for $\bar{X}$ and the joint MGF of $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$
- This is useful though in making statements about the joint distribution of $\bar{X}$ and linear combinations of the elements in the vector $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$
  - such as the joint distribution of $\bar{X}$ and S^2

Joint distribution of $\bar{X}$ and S^2

Corollary

$\bar{X}$ and S^2 are independently distributed.

- This is a direct result of the previous theorem, and is due to the fact that S^2 is a function of $(X_1 - \bar{X}, X_2 - \bar{X}, \ldots, X_n - \bar{X})$.
- From here, we now know that the joint distribution of the sample mean and sample variance is the product of the marginals.
- At this point we only know the marginal distribution of the sample mean.
- But now we have all the pieces to find the marginal distribution of S^2

Distribution of the Sample Variance

Distribution of S^2

The distribution of $\frac{(n-1)S^2}{\sigma^2}$ is the chi-square distribution with $n-1$ degrees of freedom.

Proof:

Example: Bottle filling revisited

We previously assumed that amount of liquid dispensed had $N(\mu, 1)$. Suppose we select 10 bottles and measure the amount of liquid in each. If we use these 10 observations to calculate S^2, find the numbers b_1 and b_2 such that $P(b_1 \leq S^2 \leq b_2) = 0.9$.

- To find these values, we just need to make S^2 into $\frac{(n-1)S^2}{\sigma^2}$.

$$P(b_1 \leq S^2 \leq b_2) = P\left(\frac{(n-1)b_1}{\sigma^2} \leq \frac{(n-1)S^2}{\sigma^2} \leq \frac{(n-1)b_2}{\sigma^2}\right)$$

- Once we do this, we have a $\chi^2_{(n-1)} = \chi^2_9$ and can use the table to find values a_1 and a_2 such that

$$P(a_1 \leq \chi^2_9 \leq a_2) = 0.9$$

- We can approach this like we did earlier.

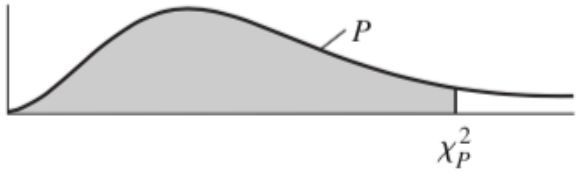
Example: Bottle filling revisited (cont.)

- To have 0.9 probability of being between two values is equivalent to having 0.1 probability being outside the interval.
- But we can be outside the interval by being either $x > a_2$ and also $x < a_1$, so we divide this 0.1 probability over the two areas.
- Therefore we have 0.05 probability of being in each tail of the chi-square distribution.
- So since the chi-square table only gives us probabilities of the form $P(X \leq x)$, we can find

$$0.9 = P(a_1 \leq \chi^2_9 \leq a_2) = P(\chi^2_9 \leq a_2) - P(\chi^2_9 \leq a_1) = 0.95 - 0.05$$

Chi-square Table of Percentiles

TABLE 3 Percentiles of the χ^2 Distribution—Values of χ^2_P Corresponding to P



df	$\chi^2_{.005}$	$\chi^2_{.01}$	$\chi^2_{.025}$	$\chi^2_{.05}$	$\chi^2_{.10}$	$\chi^2_{.90}$	$\chi^2_{.95}$	$\chi^2_{.975}$	$\chi^2_{.99}$	$\chi^2_{.995}$
1	.000039	.00016	.00098	.0039	.0158	2.71	3.84	5.02	6.63	7.88
2	.0100	.0201	.0506	.1026	.2107	4.61	5.99	7.38	9.21	10.60
3	.0717	.115	.216	.352	.584	6.25	7.81	9.35	11.34	12.84
4	.207	.297	.484	.711	1.064	7.78	9.49	11.14	13.28	14.86
5	.412	.554	.831	1.15	1.61	9.24	11.07	12.83	15.09	16.75
6	.676	.872	1.24	1.64	2.20	10.64	12.59	14.45	16.81	18.55
7	.989	1.24	1.69	2.17	2.83	12.02	14.07	16.01	18.48	20.28
8	1.34	1.65	2.18	2.73	3.49	13.36	15.51	17.53	20.09	21.96
9	1.73	2.09	2.70	3.33	4.17	14.68	16.92	19.02	21.67	23.59
10	2.16	2.56	3.25	3.94	4.87	15.99	18.31	20.48	23.21	25.19

$$a_1 = 3.33 = \frac{(n-1)b_1}{\sigma^2} = 9b_1 \quad \Rightarrow b_1 = 0.37$$

$$a_2 = 16.92 = \frac{(n-1)b_2}{\sigma^2} = 9b_2 \quad \Rightarrow b_2 = 1.88$$

Exercise - Give it a try!

Ammeters produced by a manufacturer are said to be guaranteed to give readings with a standard deviation no larger than 0.2 amps. One such meter was used to make 10 independent readings. If the sample variance of these measurements was 0.065, and readings are Normally distributed, does it suggest that this meter does not meet the manufacturer guarantee? *(Hint: find* $P(S^2 \geq 0.065)$*)*

Inference about the Mean

- Our final distributional result using the sample mean and variance is an important one for statistical inference.
- It is often of interest to want to determine where the centre of a population distribution is, or specifically if it is located in a specific place (i.e. μ equals a certain value)
- To do this, we collect a sample of data and use the sample mean as an estimate for the true mean.
- Since to use the CLT and other results, we usually need to know the variance in the population σ^2 as well.
- In practice however, we don't usually know σ^2 and therefore must also estimate it from our sample.

When the CLT doesn't quite apply

- Recall that we found the following results:
  - $\bar{X} \sim N(\mu, \sigma^2/n) \Rightarrow \bar{X} - \mu \sim N(0, \sigma^2/n) \Rightarrow \frac{\bar{X}-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$
  - $\frac{(n-1)S^2}{\sigma^2} \sim \chi^2_{(n-1)}$
- If we don't know the population variance σ^2, we have to estimate it using S^2.
  - So we would instead have $\frac{\bar{X}-\mu}{S/\sqrt{n}} = \sqrt{n}\left(\frac{\bar{X}-\mu}{S}\right)$
- But the CLT only works using σ not S
- This is because σ is a constant, but S is a random variable and thus the value that we can observe will vary from sample to sample.
  - The CLT and subsequently the standard Normal distribution does not account for this extra variability.
  - But there is a distributional result that does.

Inferences about the Mean

Corollary

Let $\bar{X}$ and S^2 be as given. Then

$$\frac{\bar{X} - \mu}{S/\sqrt{n}}$$

will have a T distribution with $n - 1$ degrees of freedom.

Proof:

Example: Forester and Fertilization

A forester is studying effects of fertilization on the average basal area of pine trees. Such basal areas are normally distributed but the forester is uncertain about the variance and thus will use his sample to estimate it. If a random sample of 9 basal areas is to be measured, find the statistics g_1 and g_2 such that $P(g_1 \leq (\bar{Y} - \mu) \leq g_2) = 0.90$.

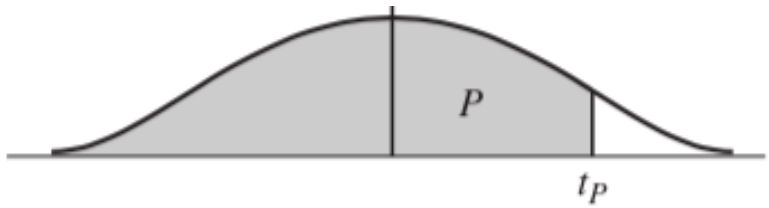
- Based on our previous result, we know that since σ is unknown we cannot use CLT, but can use the T distribution to find these values.
- Therefore, we need to rewrite the question so it uses the T distribution:

$$P\left(\frac{g_1}{S/\sqrt{n}} \leq \frac{\bar{Y} - \mu}{S/\sqrt{n}} \leq \frac{g_2}{S/\sqrt{n}}\right) = P(a_1 \leq T \leq a_2) = 0.9$$

Example: Forester and Fertilization (cont.)

- By the same logic as before, we can rewrite this in terms of values the T table will provide to us:

$$P(a_1 \leq T_{(n-1)} \leq a_2) = P(T_9 \leq a_2) - P(T_9 \leq a_1) = 0.95 - 0.05$$



df	$t_{.60}$	$t_{.70}$	$t_{.80}$	$t_{.90}$	$t_{.95}$	$t_{.975}$	$t_{.99}$	$t_{.995}$
1	.325	.727	1.376	3.078	6.314	12.706	31.821	63.657
2	.289	.617	1.061	1.886	2.920	4.303	6.965	9.925
3	.277	.584	.978	1.638	2.353	3.182	4.541	5.841
4	.271	.569	.941	1.533	2.132	2.776	3.747	4.604
5	.267	.559	.920	1.476	2.015	2.571	3.365	4.032
6	.265	.553	.906	1.440	1.943	2.447	3.143	3.707
7	.263	.549	.896	1.415	1.895	2.365	2.998	3.499
8	.262	.546	.889	1.397	1.860	2.306	2.896	3.355
9	.261	.543	.883	1.383	1.833	2.262	2.821	3.250
10	.260	.542	.879	1.372	1.812	2.228	2.764	3.169

Example: Forester and Fertilization (cont.)

- From the table we find:

$$a_2 = 1.833 = \frac{g_2}{S/\sqrt{9}} \Rightarrow g_2 = 0.611S$$

and to get the other value, use the fact that the T distribution is symmetric around 0, so

$$a_1 = -a_2 = -1.833 \Rightarrow g_1 = -0.611S$$

- Therefore, g_1 and g_2 are statistics because they rely on us being able estimate S from our data, and will thus change depending on the sample we collect.

Concluding Remarks

- All of this last week of material will be covered again at the beginning of STA261.
- It is important to understand how these distributions are related to each other.
- For your final, you only need to be comfortable with working with the sample mean and standard Normal tables.

- I will have extra office hours on Monday December 9 from 12-3pm.
- No class on the Make-up Monday (Dec. 4), for LEC0101.

Happy Studying and Good Luck!