Chapter 6

Small sample inference

Prozac anyone?

It is sometimes expensive or simply impossible to collect large samples (n>50). Consider the following study...

A study on the effect of prozac (antidepressant) on 9 patients was made.

Patients were asked to rate their "well being" before and after taking a prozac.

Before	3	0	6	7	4	3	2	1	4
After	5	1	5	7	10	9	7	11	8

It is expensive to collect more observations

Ok what about Disneyland?

Disney opened its European park in Paris in 1992.

They want to compare its performance with the performance of Disneyland California in Anaheim

Absolute number of visitors (in million) are given in the following table.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
CA	11.6	11.4	10.3	14.1	15	14.2	13.7	13.5	13.9	12.3	12.7	12.7	13.3	14.26	14.73	14.87	14.29
Paris	10	9.8	8.8	10.7	11.7	12.6	12.5	12.5	12.0	12.2	10.3	10.2	10.2	10.2	10.6	12.0	12.7

Limited time frame

The numbers are not comparable so only the increase in visitors is recorded.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
CA	NA	-0.2	-1.1	3.8	0.9	-0.8	-0.5	-0.2	0.4	-1.6	0.4	0	0.6	0.96	0.47	0.14	-0.58
Paris	NA	-0.2	-1	1.9	1	0.9	-0.1	0	-0.5	0.2	-1.9	-0.1	0	0	0.4	1.4	0.7

EuroDisney opened in 1992 so clearly there is no more data available!
It is impossible to have more than 16 observations

Bye bye CLT

We still observe $X_1, \ldots, X_n,\ i.i.d\ E(X_i) = \mu, \mathrm{var}(X_i) = \sigma^2$

Let us recall how we used the Central Limit Theorem:

If n>50 then

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \quad)$$

regardless of the distribution of $X_1, \ldots, X_n$

In particular, the distribution of $X_1, \ldots, X_n$ could be Bernoulli, Poisson, Chi-square, ... anything really

Bye bye CLT

If n>50 then

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \quad)$$

regardless of the distribution of $X_1, \ldots, X_n$

The normal distribution allow<u>ed</u> us to say that the Z-score

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

Bye bye CLT

If n>50 then

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \quad)$$

regardless of the distribution of $X_1, \ldots, X_n$

The normal distribution allowed us to say that the Z-score

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0, 1) \text{ approximately}$$

Bye bye CLT

If n>50 then

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \quad)$$

regardless of the distribution of $X_1, \ldots, X_n$

The normal distribution allowed us to say that the Z-score

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1) \text{ approximately}$$

This enabled us to use the table for the standard normal distribution $\longrightarrow$ confidence intervals, p-values

Bye bye CLT

If n>50 then

$$\bar{X} = \frac{X_1 + \ldots + X_n}{n} \sim N(\mu, \quad)$$

regardless of the distribution of $X_1, \ldots, X_n$

The normal distribution allow<u>ed</u> us to say that the Z-score

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim N(0,1) \text{ approximately}$$

This enabl<u>ed</u> us to use the table for the standard normal distribution $\longrightarrow$ confidence intervals, p-values

But now n is much smaller than 50

Distribution of the Z-score

The purpose of this chapter is to find the distribution of the Z-score

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

under some assumptions even when n is small

We need to make the assumption that

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \quad i.i.d$$

This assumption should be checked with a normal QQplot!

Unknown variance

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \quad i.i.d$$

But don't we already know the distribution of the Z-score under this assumption?

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

NO! what we know is that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0,1)$$

What's the difference?

Unknown variance

$$X_1, \ldots, X_n \sim N(\mu, \sigma^2) \quad i.i.d$$

But don't we already know the distribution of the Z-score under this assumption?

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}}$$

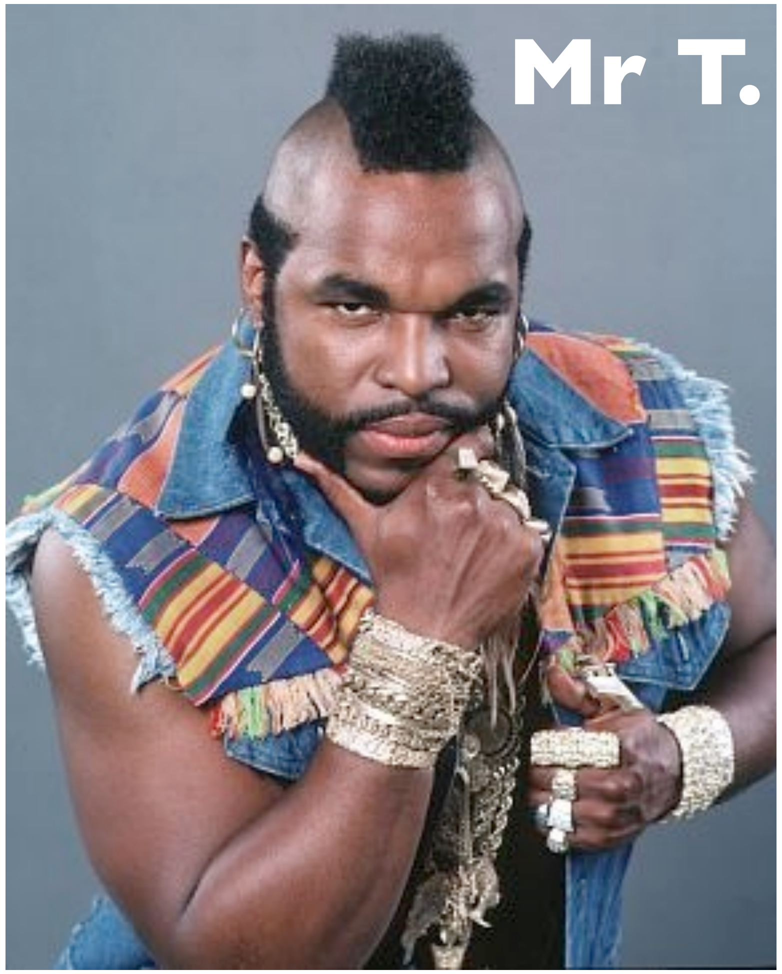
NO! what we know is that

$$Z = \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \sim N(0, 1)$$

What's the difference?

The variance σ^2 is unknown and replaced by its estimator s^2

The t distribution



RULE

If $X_1, \ldots, X_n \sim N(\mu, \sigma^2) \quad i.i.d$

Then $$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_{n-1}$$

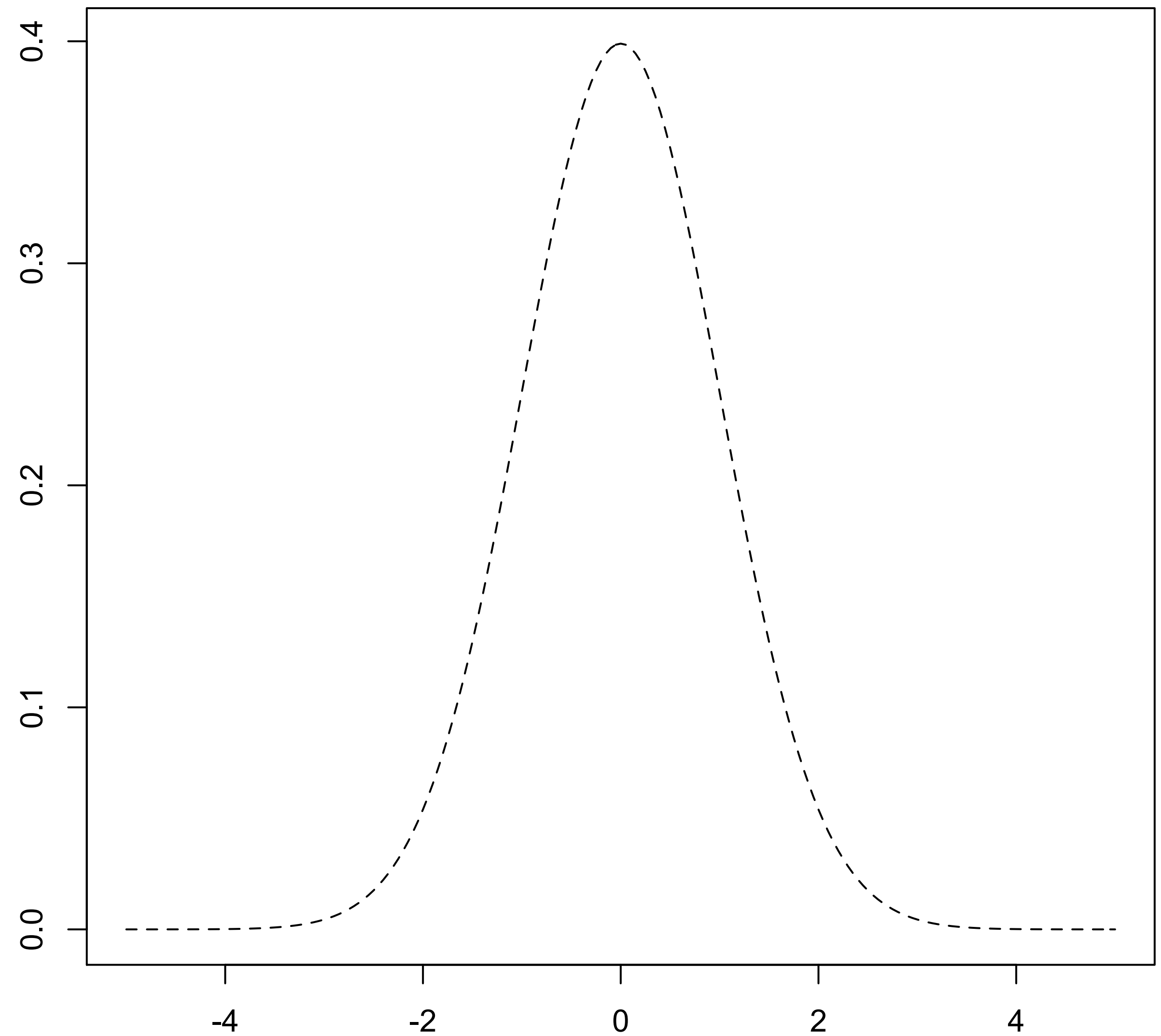
This distribution is NOT the standard normal distribution.

It has one integer parameter (here n-1) called
degrees of freedom (d.f.)

The t distribution

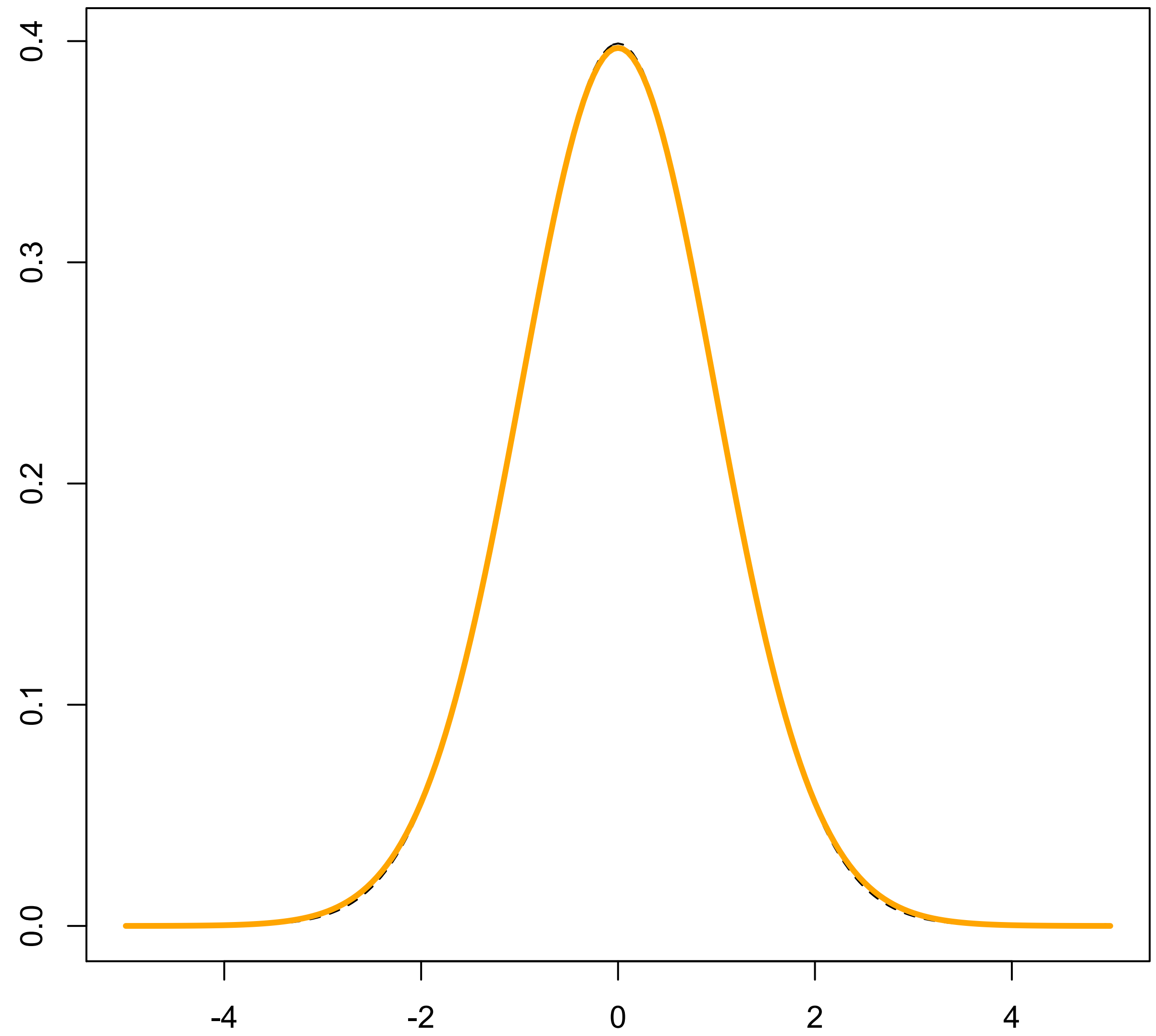
Actually this distribution was used first by Sean William Gosset in 1908 while he worked for the Guinness brewery in Dublin Ireland. His employer forbid him to publish papers so he used the pseudo "student"

The t distribution Vs Standard normal



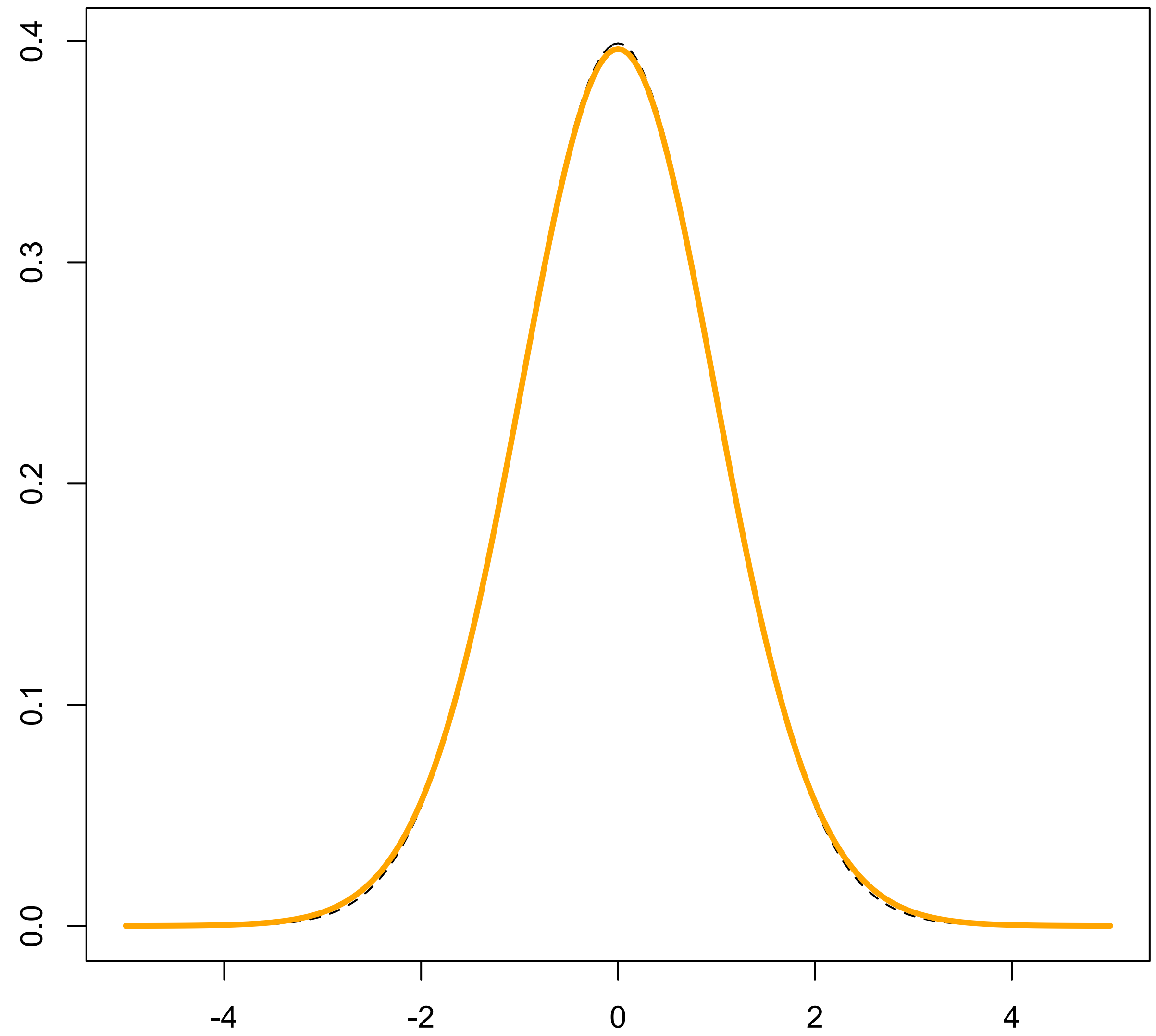
The standard normal pdf

The t distribution: df=50



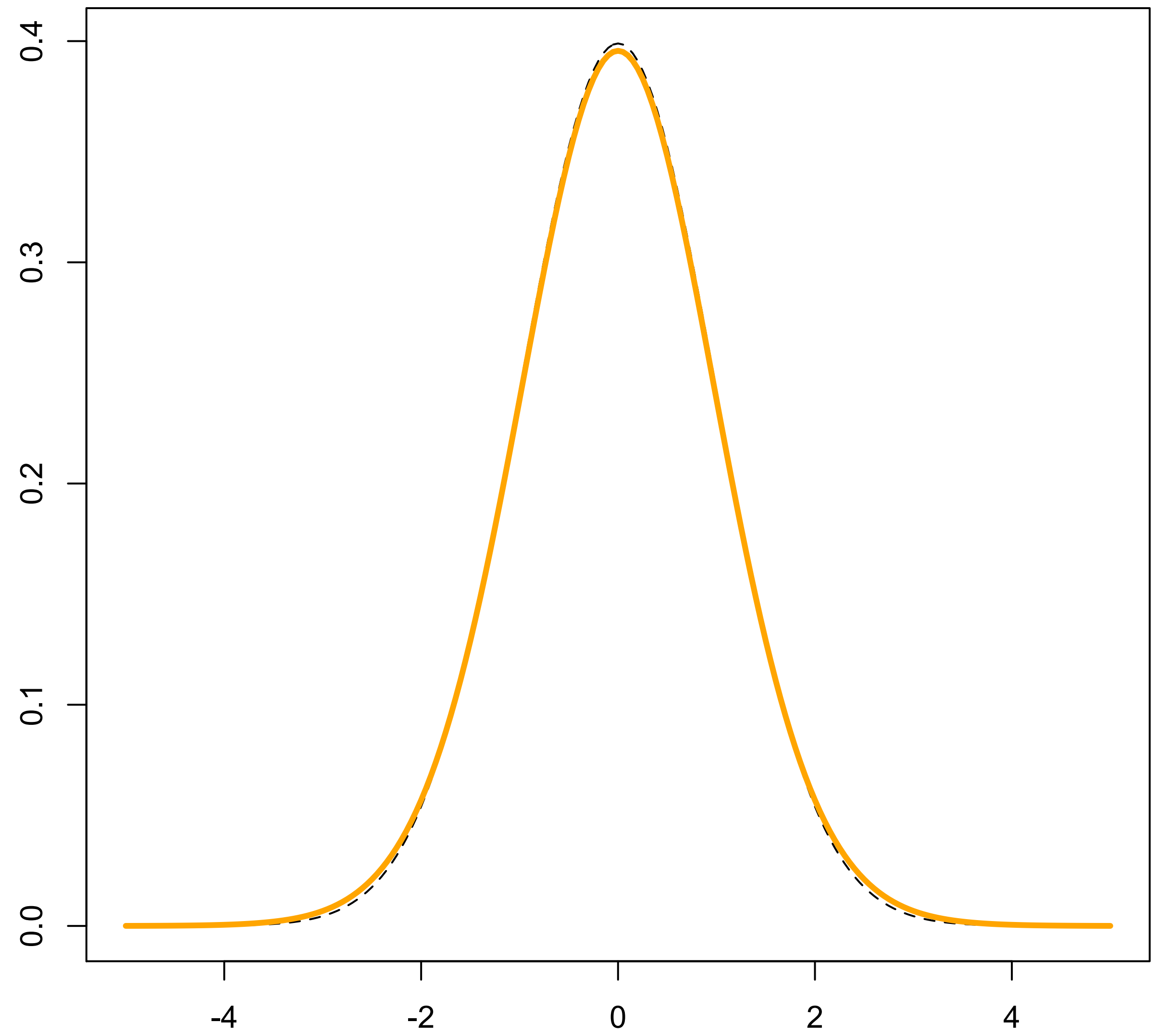
The t_{50} pdf

The t distribution: df=40



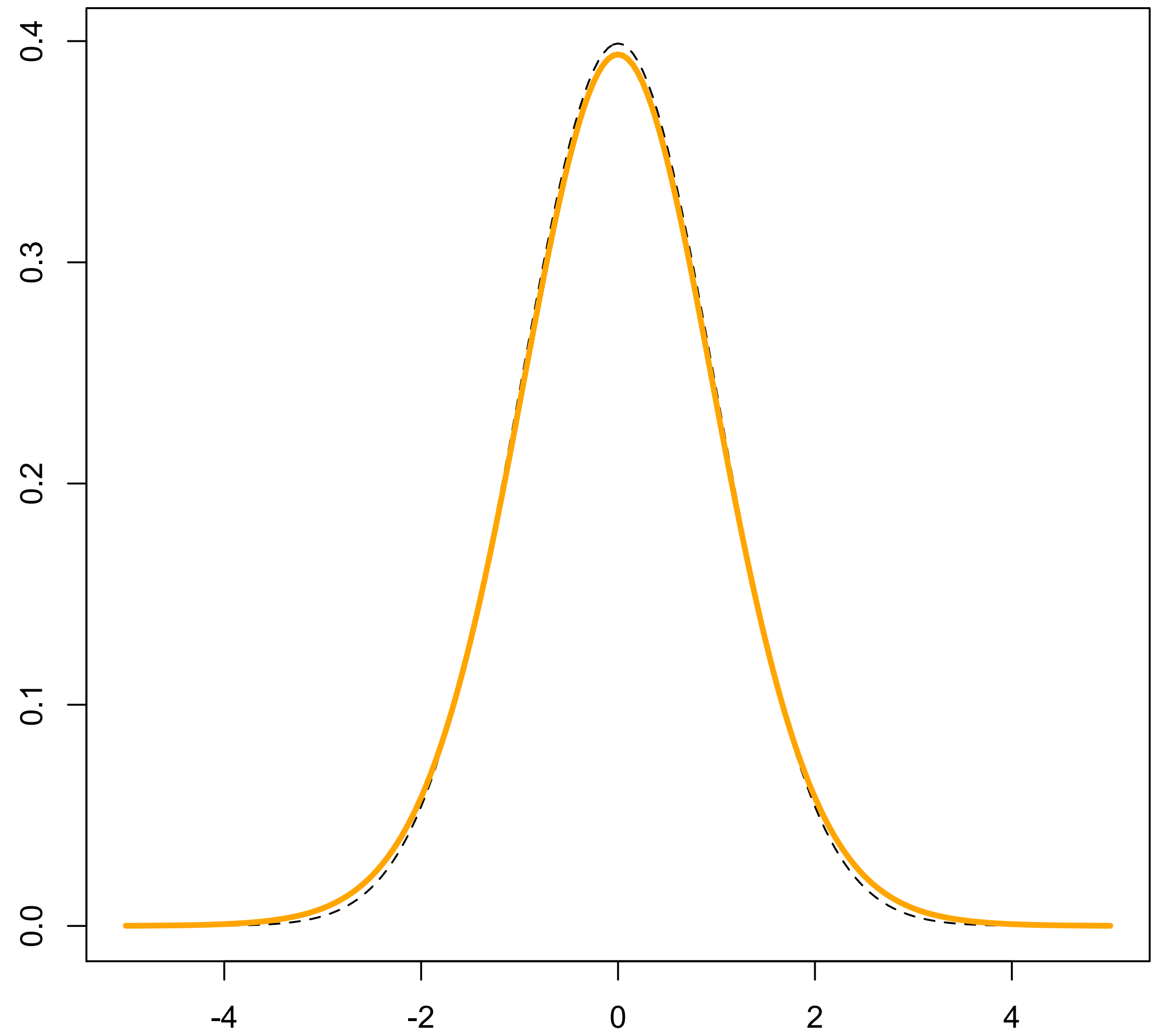
The t_{40} pdf

The t distribution: df=30



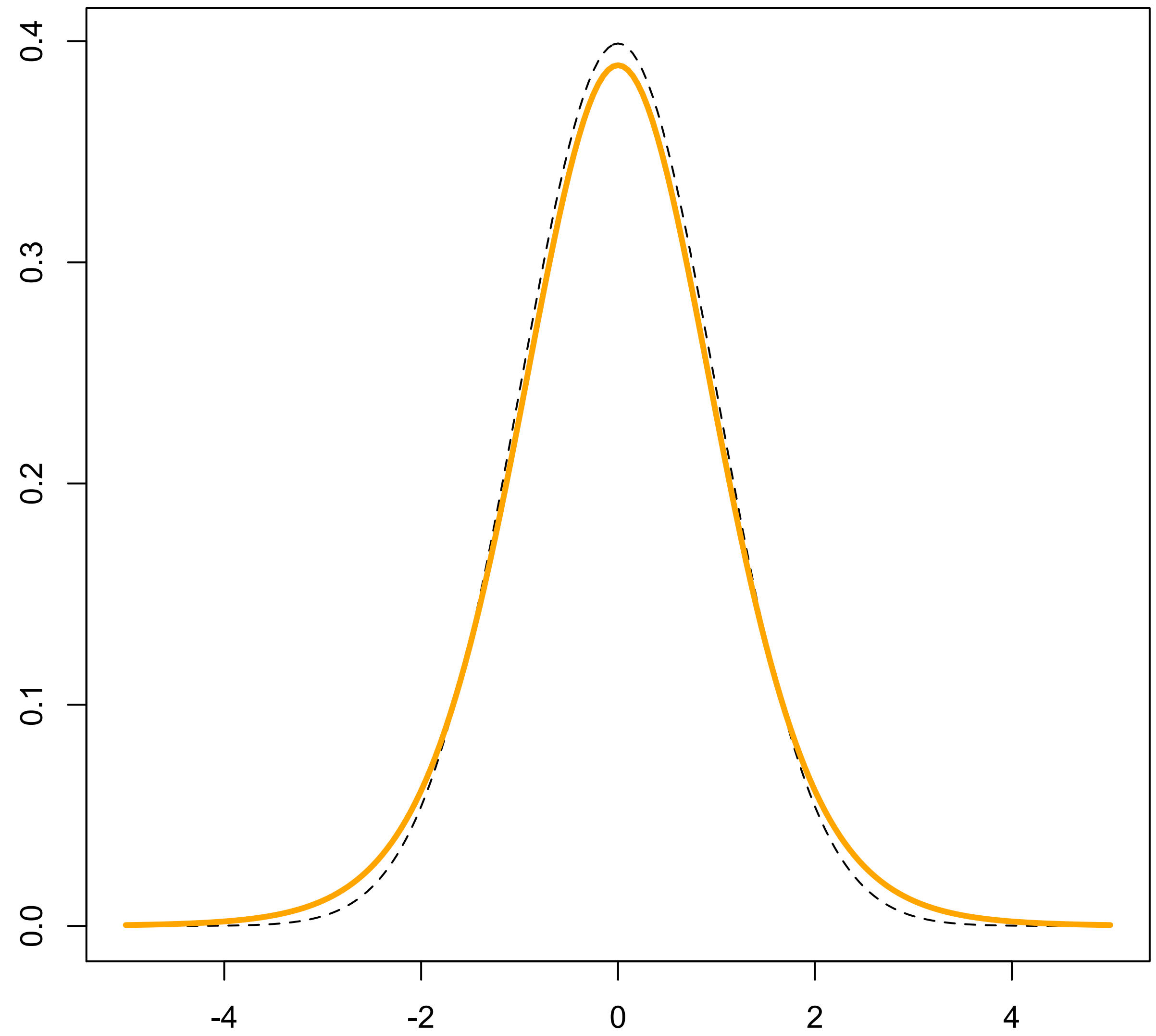
The t_{30} pdf

The t distribution: df=20



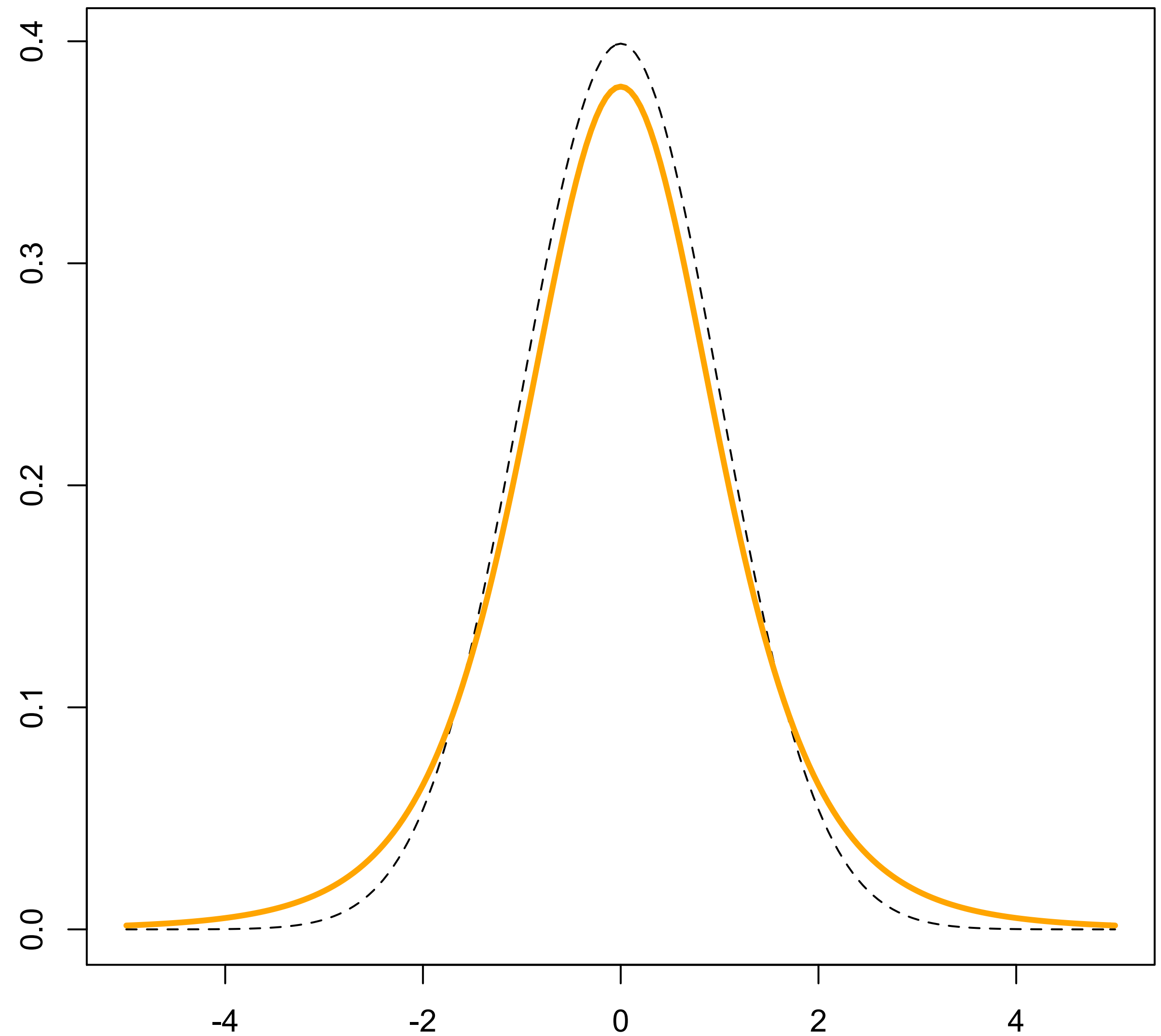
The t_{20} pdf

The t distribution: df=10



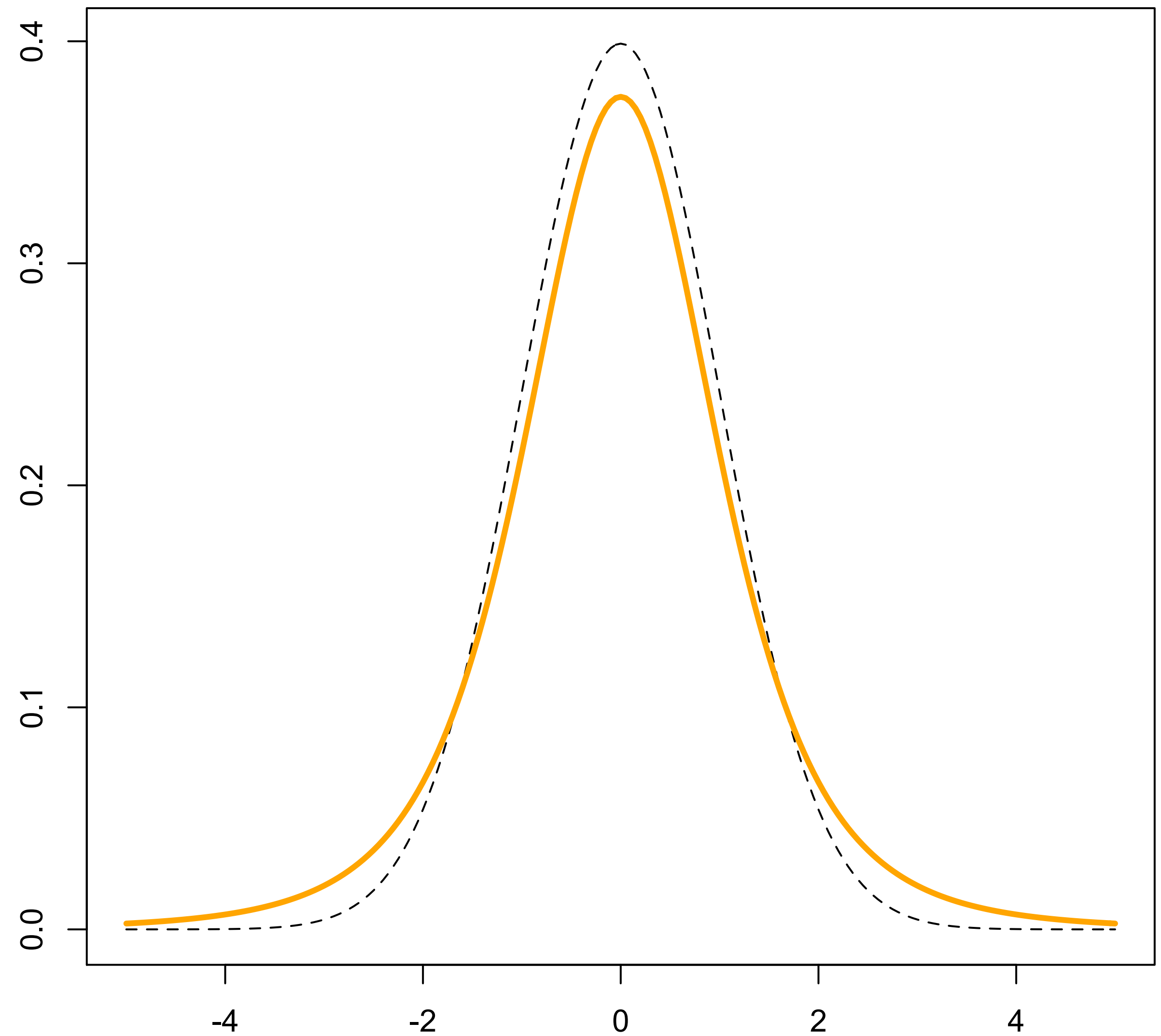
The t_{10} pdf

The t distribution: df=5



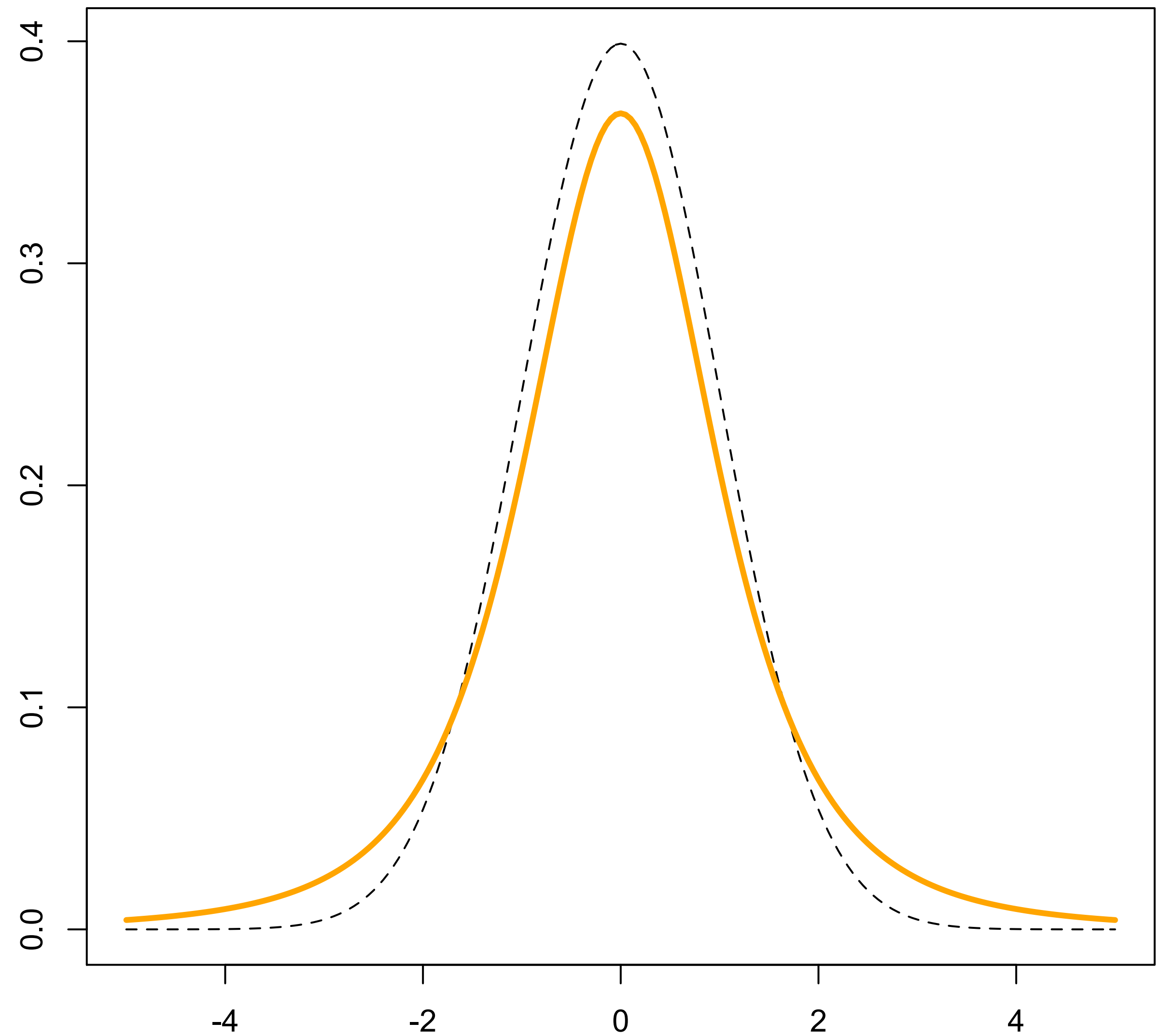
The t_5 pdf

The t distribution: df=4



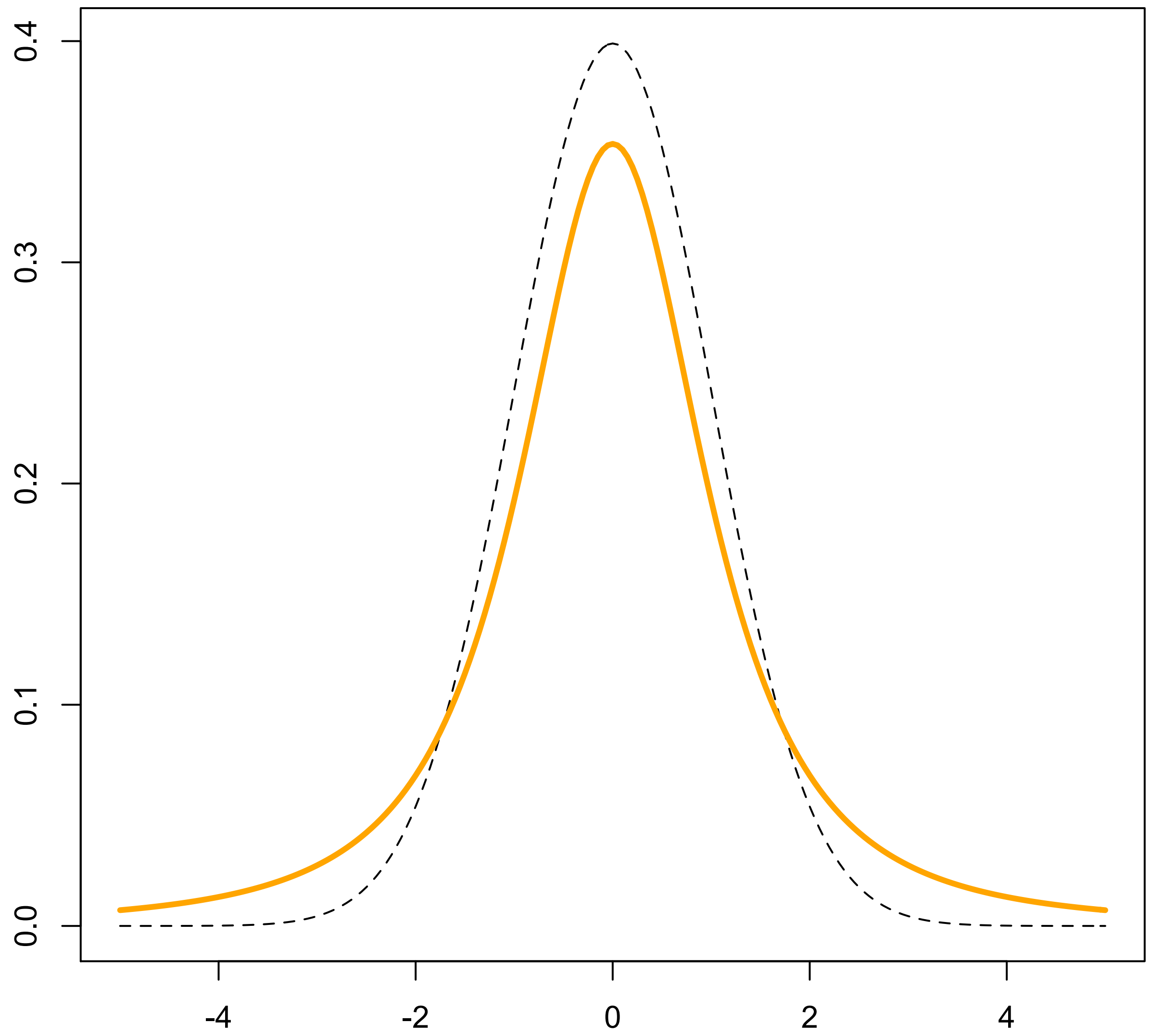
The t_4 pdf

The t distribution: df=3



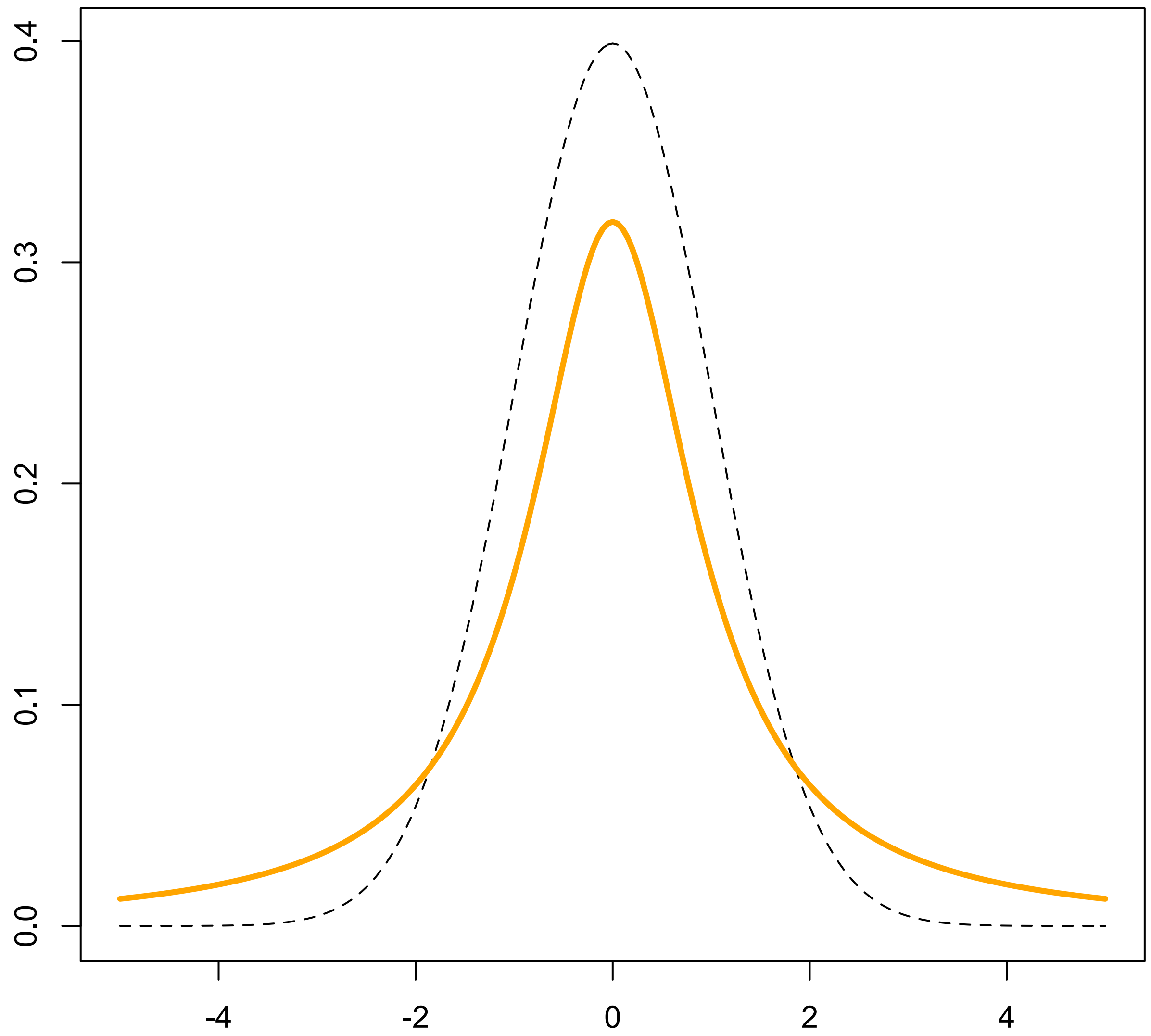
The t_3 pdf

The t distribution: df=2



The t_2 pdf

The t distribution: df=1



The t_1 pdf

The t distribution `with R`

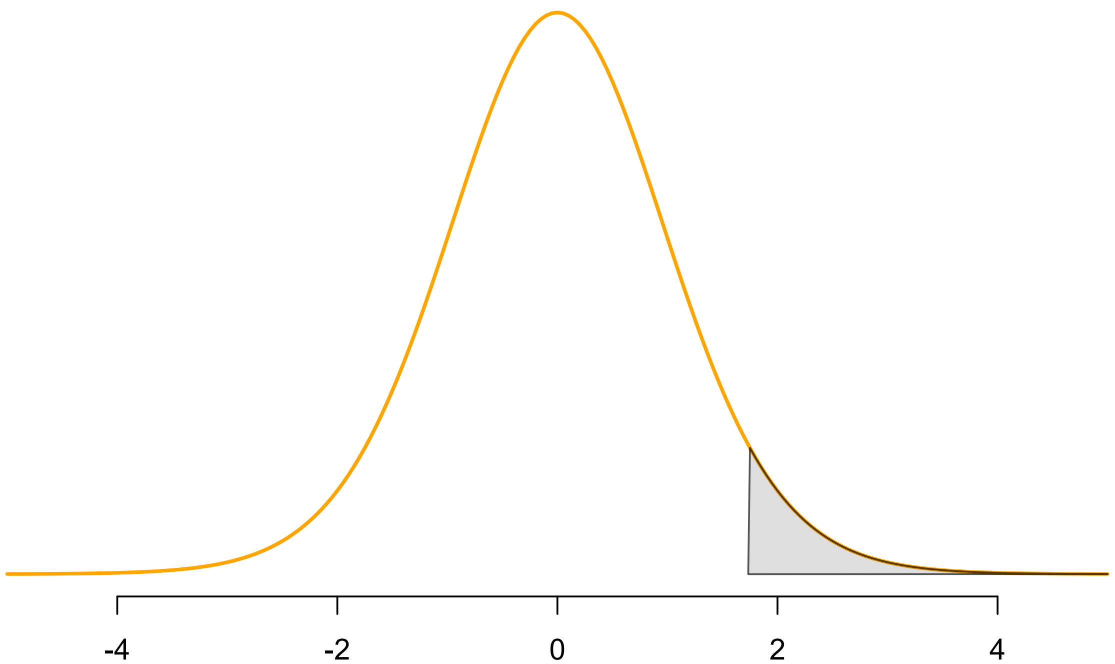
	Normal	t_{13}
Random sample	`rnorm(100)`	`rt(100,df=13)`
pdf	`dnorm(x)`	`dt(x,df=13)`
quantiles	`qnorm(x)`	`qt(x,df=13)`

The t distribution by hand

There are also tables for the t distribution: one table per d.f. (just like the chi-square). Here is an abbreviated version:

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	*1.33*	*1.73*	*2.10*	*2.55*	*2.88*
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85
	⋮	⋮	⋮	⋮	⋮	
	400	1.28	1.65	1.97	2.34	2.59
	500	1.28	1.65	1.96	2.33	2.59
	∞	1.28	1.64	1.96	2.33	2.58

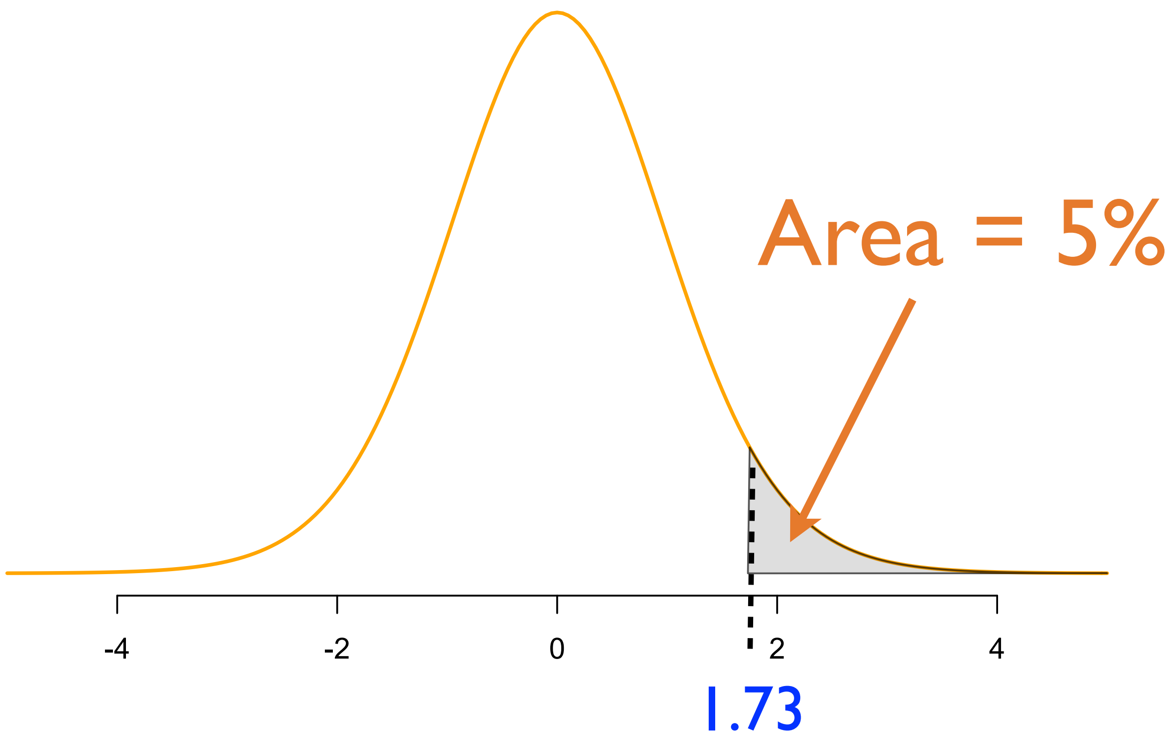
What am I reading?

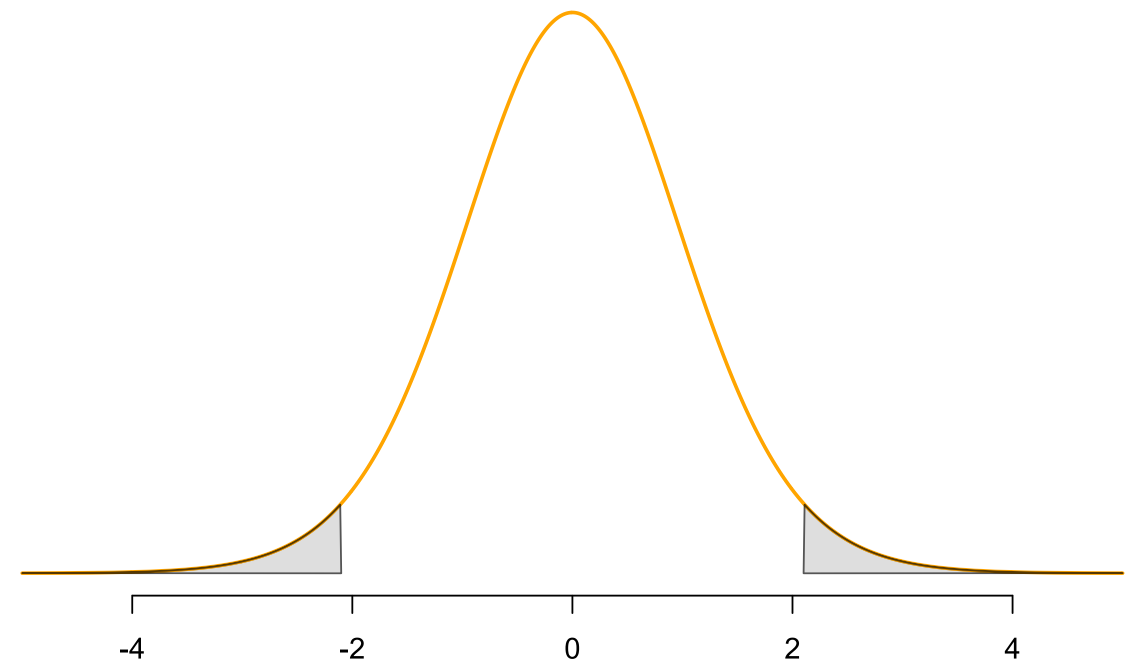


	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	*1.33*	*1.73*	*2.10*	*2.55*	*2.88*
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

```
> x=seq(-5, 5, by=0.05)
> plot(x, dt(x), lwd=2, col='orange', xlab="", ylab="",  axes=FALSE, type="l")
> z=qt(0.95, df=18)
polygon(c(x[x>z],z),c(dt(x[x>z], df=18),0), col='#00000022', border='#000000AA')
```

What am I reading?

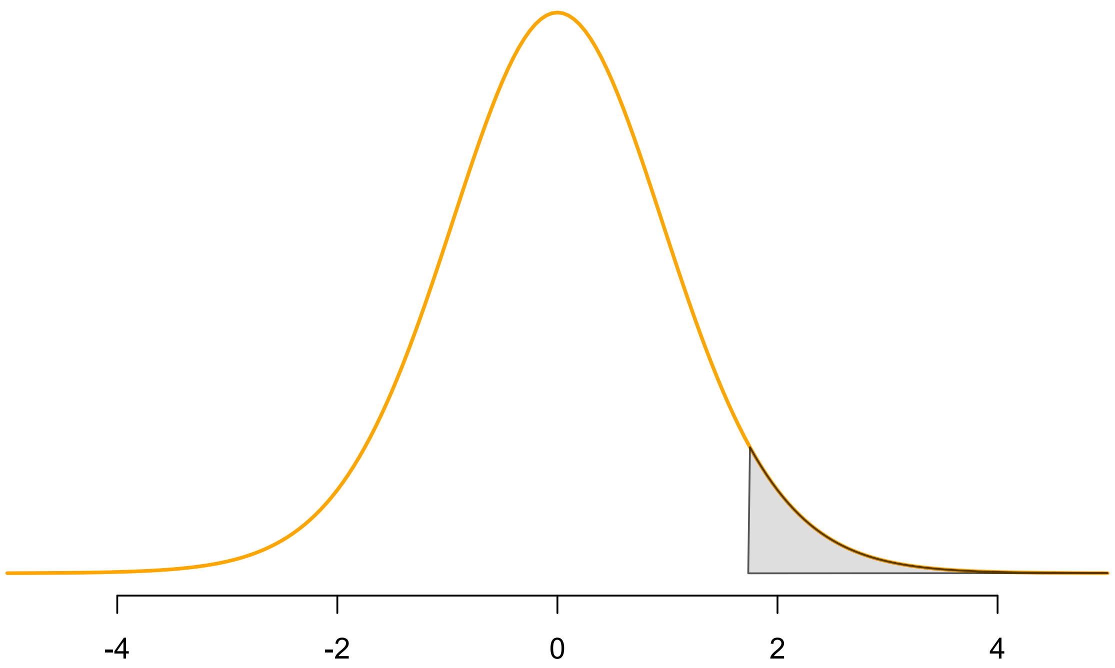


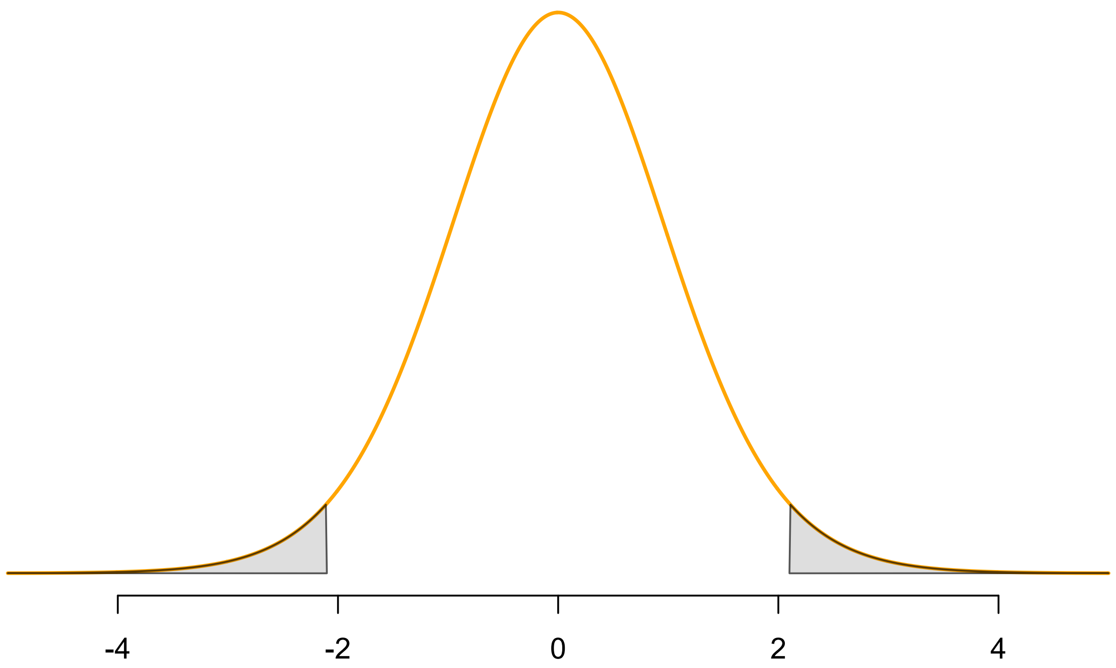


one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	*1.33*	*1.73*	*2.10*	*2.55*	*2.88*
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

```
> x=seq(-5, 5, by=0.05)
> plot(x, dt(x), lwd=2, col='orange', xlab="", ylab="",  axes=FALSE, type="l")
> z=qt(0.95, df=18)
polygon(c(x[x>z],z),c(dt(x[x>z], df=18),0), col='#00000022', border='#000000AA')
```

What am I reading?

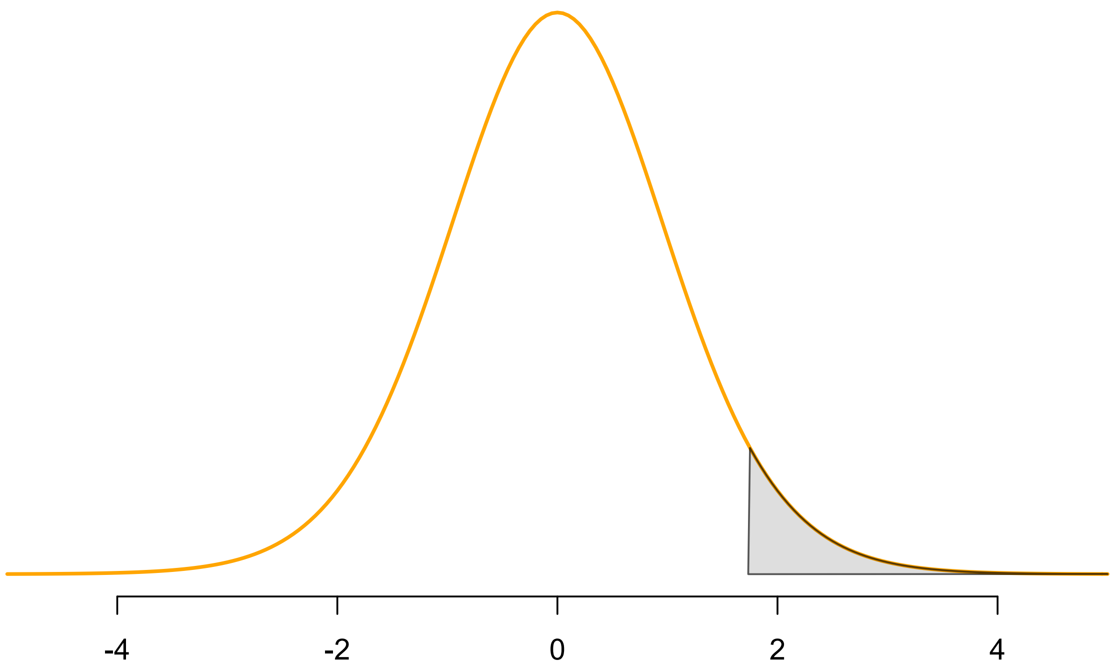


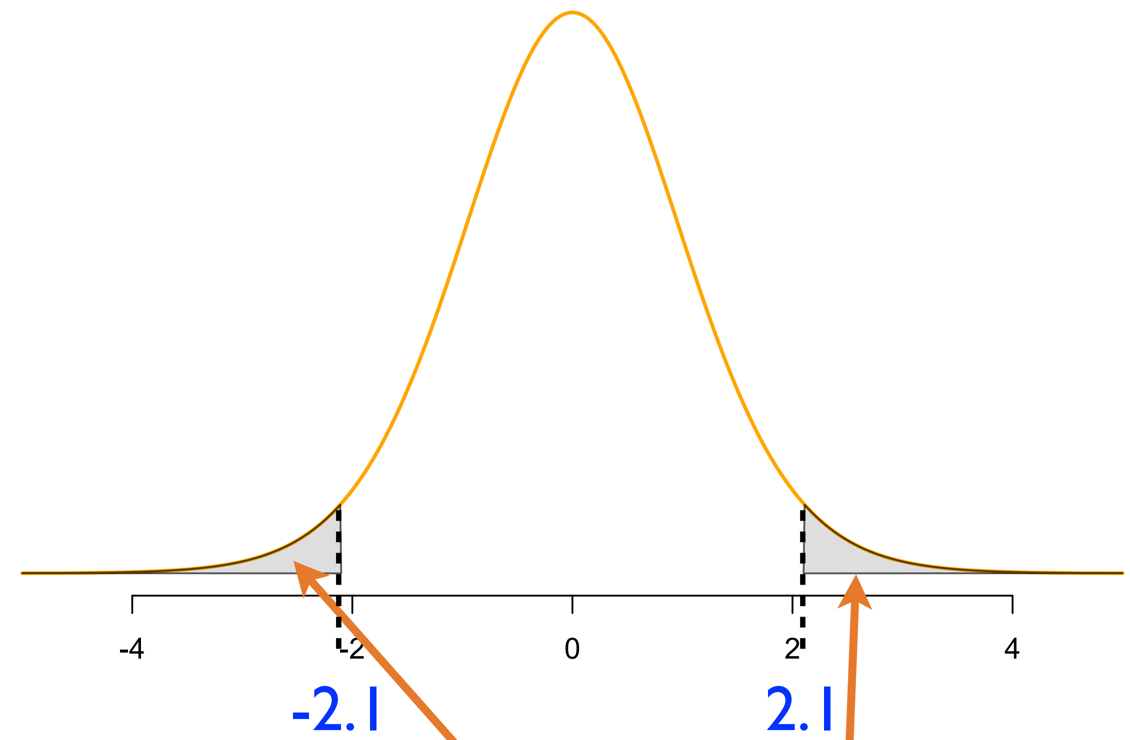


one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	*1.33*	*1.73*	*2.10*	*2.55*	*2.88*
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

```
> x=seq(-5, 5, by=0.05)
> plot(x, dt(x), lwd=2, col='orange', xlab="", ylab="",  axes=FALSE, type="l")
> z=qt(0.95, df=18)
polygon(c(x[x>z],z),c(dt(x[x>z], df=18),0), col='#00000022', border='#000000AA')
```

What am I reading?





Area = 2.5+2.5=5%

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	⋮	⋮	⋮	⋮	⋮	
	17	1.33	1.74	2.11	2.57	2.90
	18	*1.33*	*1.73*	*2.10*	*2.55*	*2.88*
	19	1.33	1.73	2.09	2.54	2.86
	20	1.33	1.72	2.09	2.53	2.85

```
> x=seq(-5, 5, by=0.05)
> plot(x, dt(x), lwd=2, col='orange', xlab="", ylab="",  axes=FALSE, type="l")
> z=qt(0.95, df=18)
polygon(c(x[x>z],z),c(dt(x[x>z], df=18),0), col='#00000022', border='#000000AA')
```

What can we do now?

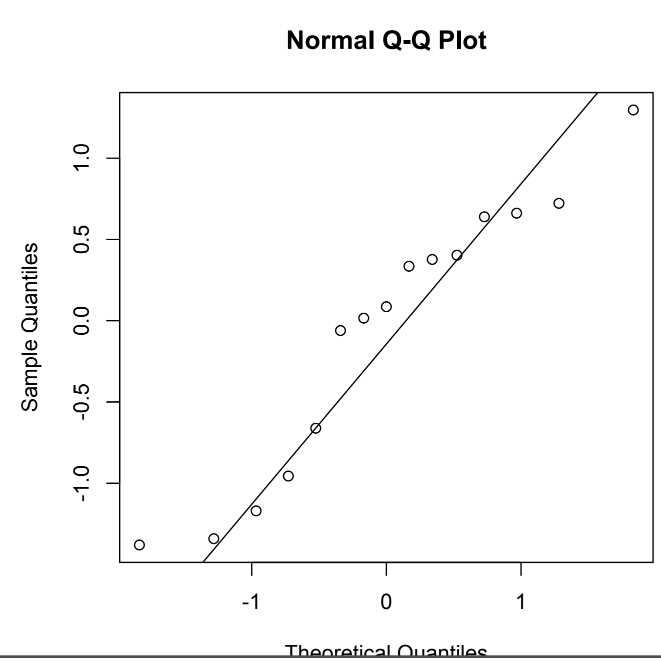
We can do the same things as in the large sample case:

1. Confidence interval for the mean μ
2. Test for the mean μ
3. Confidence interval for the difference between means $\mu_1 - \mu_2$
4. Test for the difference between means $\mu_1 - \mu_2$

When can we do it?

When the observations are

approximately normal and independent



from the experiment
(there is no test for that)

Test for a mean

Before	3	0	6	7	4	3	2	1	4
After	5	1	5	7	10	9	7	11	8

We are interested in finding out if prozac actually has an effect on the mood

$$H_0 : \mu_{after} \leq \mu_{before}$$
$$H_A : \mu_{after} > \mu_{before}$$

This is paired data so we can form the difference and make a test on $\mu_d = \mu_{after} - \mu_{before}$

Test for a mean

Before	3	0	6	7	4	3	2	1	4
After	5	1	5	7	10	9	7	11	8
Diff	2	1	-1	0	6	6	5	10	4

We are interested in finding out if prozac actually has an effect on the mood

$$H_0 : \mu_{after} \leq \mu_{before}$$
$$H_A : \mu_{after} > \mu_{before}$$

This is paired data so we can form the difference and make a test on $\mu_d = \mu_{after} - \mu_{before}$

It is the same thing as testing a performing a test for the mean (point 2.)

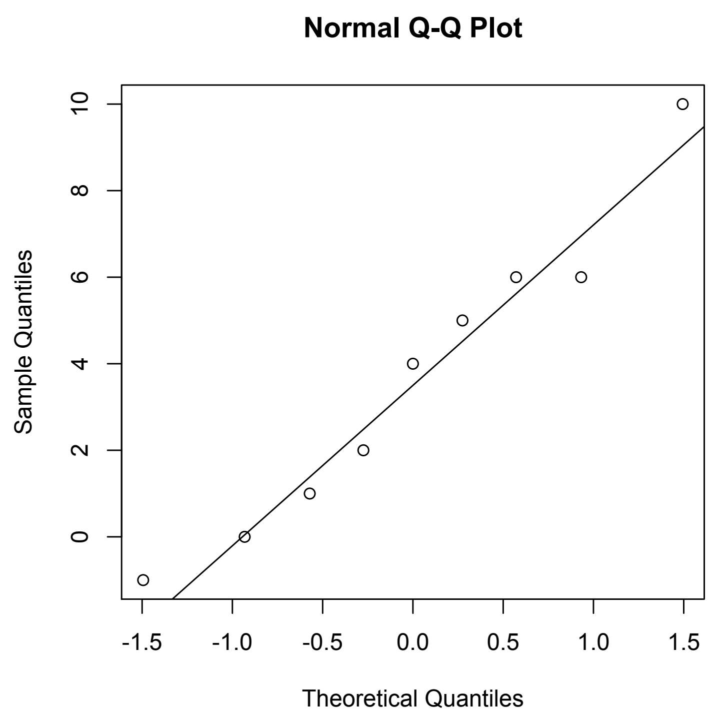
Test for a mean

Diff	2	1	-1	0	6	6	5	10	4

The test becomes

$$H_0 : \mu_d \leq 0$$
$$H_A : \mu_d > 0$$

There are only 9 observations so we cannot use the CLT.
We need to use the t distribution: **t test**
Can it be used? We need to check normality...



YES! (normal distribution)

Test for a mean

Therefore the Z-score has t distribution with 9-1=8 observations

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_8 \longrightarrow Z = \frac{\bar{X} - 0}{s/\sqrt{n}} \sim t_8$$ under H_0

If we compute the observed z-score under H_0, we find

$$z_{obs} = \frac{\bar{x} - 0}{s/\sqrt{n}} = \frac{3.67 - 0}{3.5/\sqrt{9}} = 3.14$$

Therefore the p-value is given by (one sided test):

$$\text{p-value} = P(Z > z_{obs}) = P(t_8 > 3.14)$$

$$\text{p-value} = P(Z > z_{obs}) = P(t_8 > 3.14)$$

	one tail	0.100	0.050	0.025	0.010	0.005
	two tails	0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36

We look at the table and find that the p-value is between 0.5% and 1%. So the conclusion is that we **reject the null hypothesis**

Prozac works!

Test for a mean `with R`

We already know how to use the test with student distribution (we've been using it all along):

```
> diff=c(2,1,-1,0,6,6,5,10,4)
> t.test(diff, alternative="greater")
```

```
        One Sample t-test

data:  diff
t = 3.1429, df = 8, p-value = 0.006873
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 1.497194       Inf
sample estimates:
mean of x
 3.666667
```

Confidence interval for the mean

We saw that R can give one sided confidence interval for the mean μ_d

We can also make two-sided confidence intervals

```
> diff=c(2,1,-1,0,6,6,5,10,4)
> t.test(diff)
```

```
        One Sample t-test

data:  diff
t = 3.1429, df = 8, p-value = 0.01375
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.9763285 6.3570048
sample estimates:
mean of x
 3.666667
```

Confidence interval for the mean

We start from the distribution of the Z-score

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_8$$

From the table, we see that $P(|Z| > 2.31) = 0.05$

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	1	3.08	6.31	12.71	31.82	63.66
	2	1.89	2.92	4.30	6.96	9.92
	3	1.64	2.35	3.18	4.54	5.84
	4	1.53	2.13	2.78	3.75	4.60
	5	1.48	2.02	2.57	3.36	4.03
	6	1.44	1.94	2.45	3.14	3.71
	7	1.41	1.89	2.36	3.00	3.50
	8	1.40	1.86	2.31	2.90	3.36

Confidence interval for the mean

$$Z = \frac{\bar{X} - \mu}{s/\sqrt{n}} \sim t_8 \qquad P(|Z| > 2.31) = 0.05$$

Plugging the definition of Z in the probability yields

$$P(|\frac{\bar{X} - \mu}{s/\sqrt{n}}| > 2.31) = 0.05$$

$$P(\bar{X} - 2.31\frac{s}{\sqrt{n}} \leq \mu \leq \bar{X} + 2.31\frac{s}{\sqrt{n}}) = 0.05$$

It gives the confidence interval

$$[\bar{x} - 2.31\frac{s}{\sqrt{n}} \ , \ \bar{x} + 2.31\frac{s}{\sqrt{n}}] = [0.97 \ , \ 6.36]$$

Difference between two means

What if the data is **not paired?**

A laboratory analysis of calories of major hot dog brands. Researchers for Consumer Reports analyzed two types of hot dog: beef and poultry. The results are summarized below:

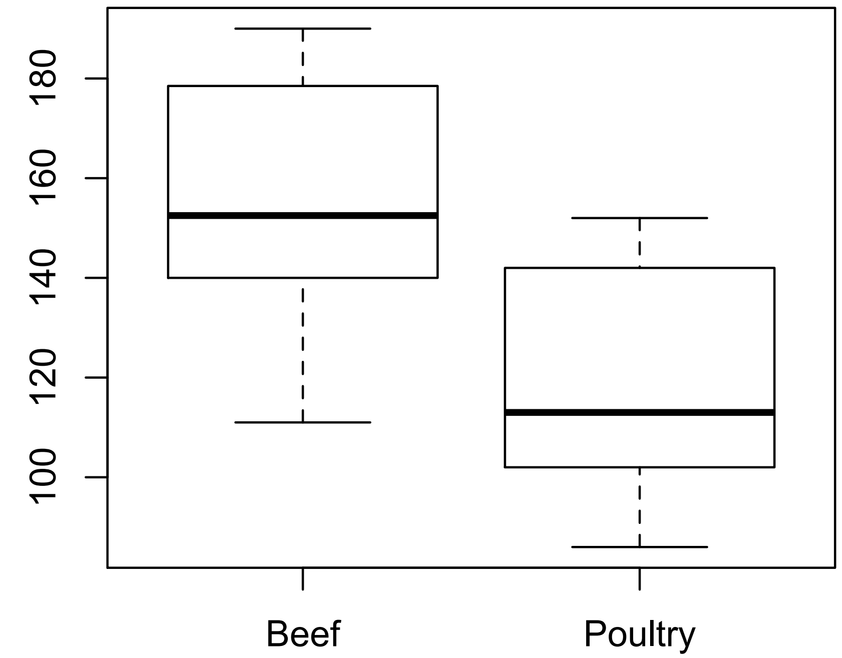
	Mean $\bar{x}$	Std-dev s	size n
 BEEF	156.85	22.64	20
 POULTRY	118.76	22.55	17

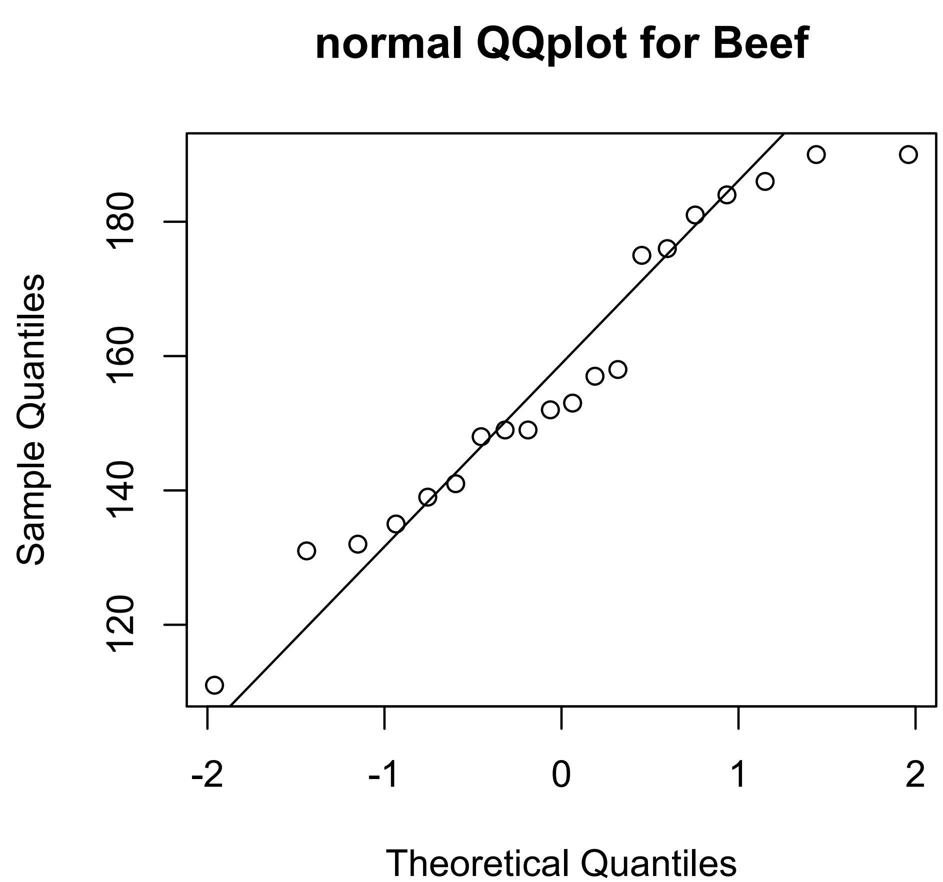
Difference between two means

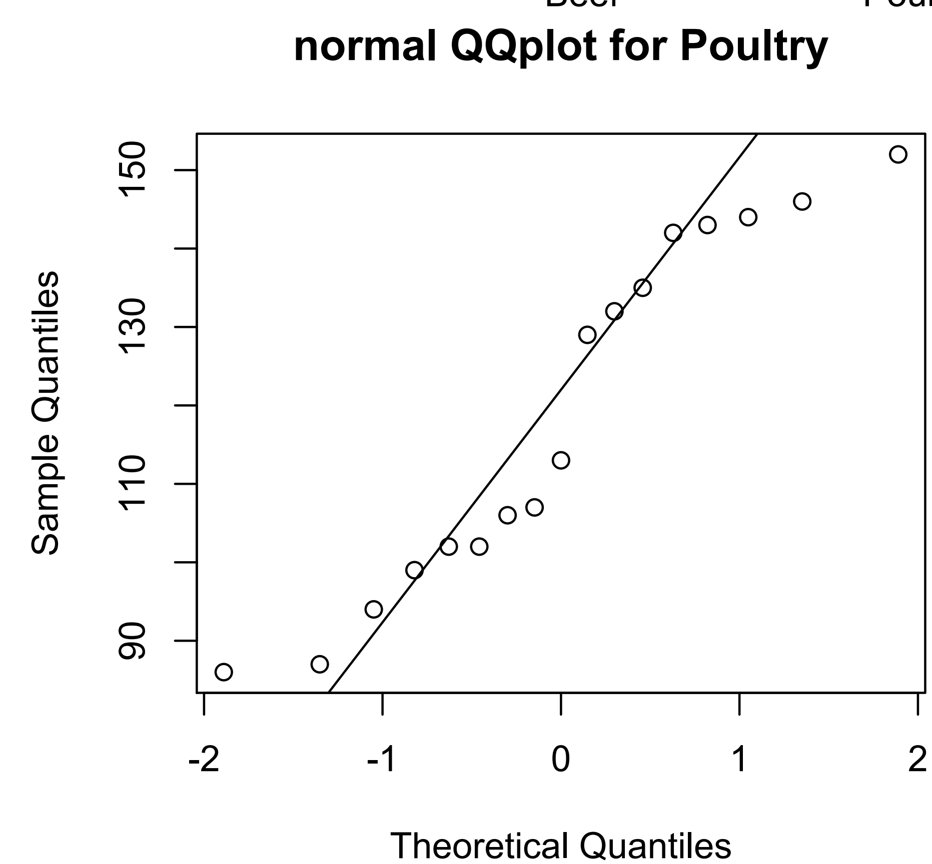
We want to know if there is a difference between beef and poultry

$$H_0 : \mu_B = \mu_P$$

$$H_A : \mu_B \neq \mu_P$$

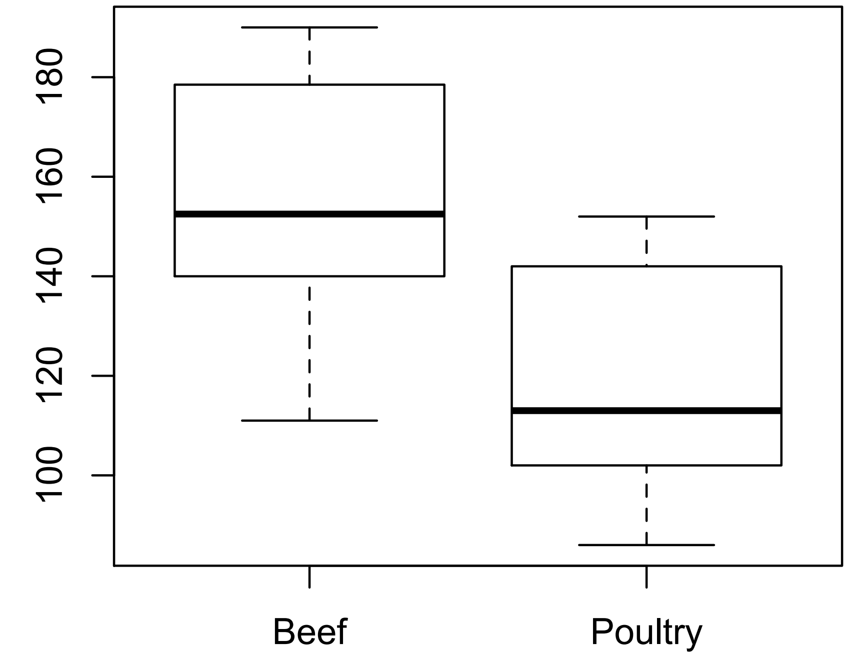






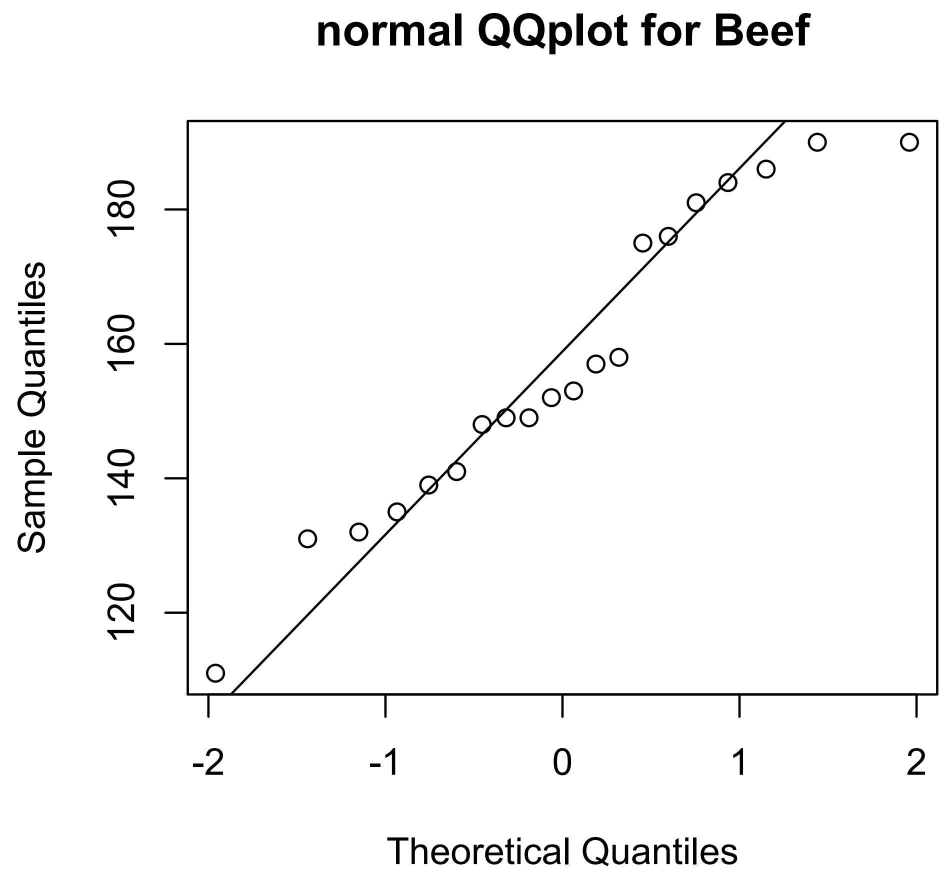
Difference between two means

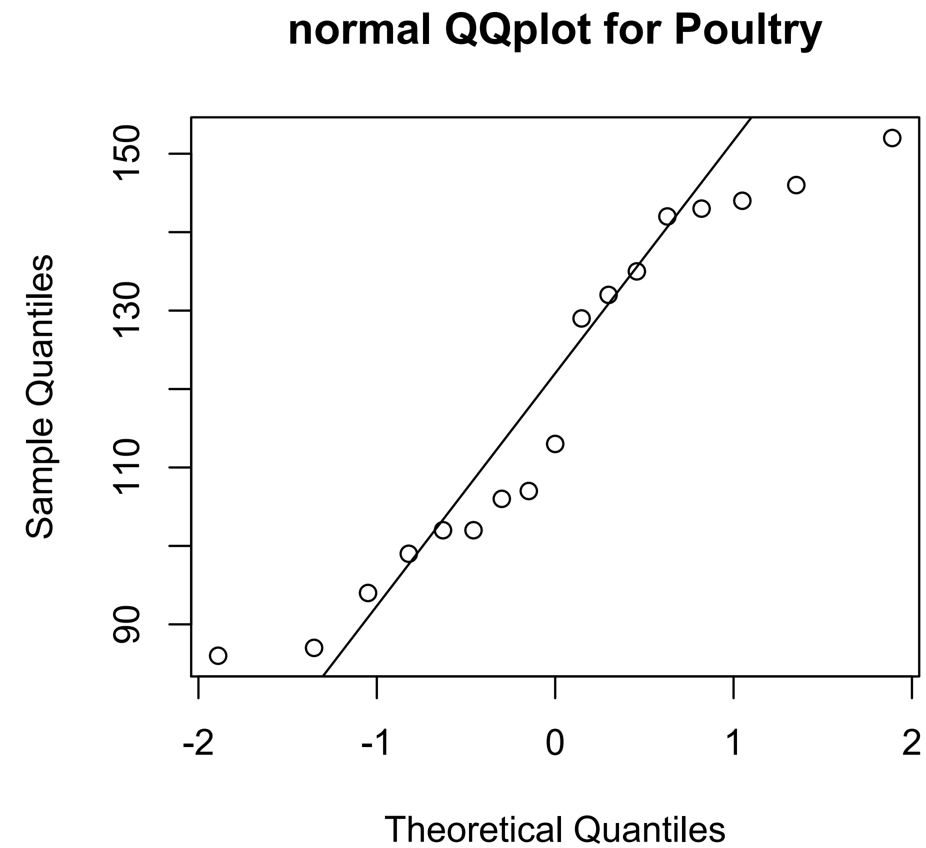
Boxplots indicate that there may be a significant difference. Can we perform a test and get a p-value?



- Sample size is too small for CLT
- We need to use the student distribution
- But the data is not paired (a hotdog is either beef or poultry)

Difference between two means





The qqplots indicate that the data is approximately normal.

If we assume that the two samples are independent, then

$$\bar{X}_B - \bar{X}_P \sim N(\qquad , \qquad) \quad \text{under} \quad H_0 : \mu_B = \mu_P$$

Difference between two means

$$\bar{X}_B - \bar{X}_P \sim N\left(0 , \frac{\sigma_B^2}{n_B} + \frac{\sigma_P^2}{n_P} \right) \quad \text{under} \quad H_0 : \mu_B = \mu_P$$

Indeed, if the sample are independent:

$$var(\bar{X}_B - \bar{X}_P) = var(\bar{X}_B) + var(\bar{X}_P)$$
$$= \frac{\sigma_B^2}{n_B} + \frac{\sigma_P^2}{n_P}$$

We form the Z-score...

Recall that we don't know the variances σ_B^2 and σ_P^2 so we replace them by s_B^2 and s_P^2 respectively!

The Z-score is

$$Z = \frac{\bar{X}_B - \bar{X}_P - (\mu_B - \mu_P)}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}}} \longrightarrow Z = \frac{\bar{X}_B - \bar{X}_P - 0}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}}} \text{ under } H_0$$

Difference between two means

$$Z = \frac{\bar{X}_B - \bar{X}_P - 0}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}}}$$

We need to find the distribution of the above Z-score

Difference between two means

$$Z = \frac{\bar{X}_B - \bar{X}_P - 0}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}}}$$

We need to find the distribution of the above Z-score

It is a t distribution

Difference between two means

$$Z = \frac{\bar{X}_B - \bar{X}_P - 0}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}}}$$

We need to find the distribution of the above Z-score

It is a t distribution

So we only need the d.f. to find which table to read from. The book says:

$$\text{df} = \min(n_B - 1, n_P - 1) = \min(20 - 1, 17 - 1) = 16$$

This is an easy rule but let's see what R does...

Difference between two means

with R

We already know how to use the test with student distribution (we've been using it all along):

```
> t.test(Beef, Poultry)
```

```
        Welch Two Sample t-test

data:  Beef and Poultry
t = 5.11, df = 34.09, p-value = 1.229e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 22.94024 53.23035
sample estimates:
mean of x mean of y
 156.8500  118.7647
```

Difference between two means

with R

We already know how to use the test with student distribution (we've been using it all along):

```
> t.test(Beef, Poultry)
```

```
Welch Two Sample t-test

data:  Beef and Poultry
t = 5.11, df = 34.09, p-value = 1.229e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 22.94024 53.23035
sample estimates:
mean of x mean of y
 156.8500  118.7647
```

we reject!

Difference between two means

with R

We already know how to use the test with student distribution (we've been using it all along):

```
> t.test(Beef, Poultry)
```

not 16!!

```
        Welch Two Sample t-test

data:  Beef and Poultry
t = 5.11, df = 34.09, p-value = 1.229e-05
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 22.94024 53.23035
sample estimates:
mean of x mean of y
 156.8500  118.7647
```

we reject!

Computing and using the df

How did R find 34.09? Complicated formula:

$$\frac{\left(\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}\right)^2}{\frac{s_B^4}{n_B^2(n_B-1)} + \frac{s_P^4}{n_P^2(n_P-1)}}$$

How can we use it with a table? We round it down (truncate)! Here we use the table for df=34 (more conservative).

P-value

R found the same value

To find the p-value, we proceed as usual.
The observed Z-score is

$$z_{obs} = \frac{\bar{x}_B - \bar{x}_P - 0}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_P^2}{n_P}}} = \frac{156.85 - 118.76}{\sqrt{\frac{22.64^2}{20} + \frac{22.55^2}{17}}} = 5.11$$

The p-value is now given by

$$\text{p-value} = P(|Z| > z_{obs}) = P(|t_{34}| > 5.11)$$

we read this value from a table

P-value

$$\text{p-value} = P(|Z| > z_{obs}) = P(|t_{34}| > 5.11)$$

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72

$$2.72 < 5.11$$

So the p-value is smaller than 1%

P-value

$$\text{p-value} = P(|Z| > z_{obs}) = P(|t_{34}| > 5.11)$$

one tail two tails	0.100 0.200	0.050 0.100	0.025 0.050	0.010 0.020	0.005 0.010
df 31	1.31	1.70	2.04	2.45	2.74
32	1.31	1.69	2.04	2.45	2.74
33	1.31	1.69	2.03	2.44	2.73
34	1.31	1.69	2.03	2.44	2.73
35	1.31	1.69	2.03	2.44	2.72

$$2.72 < 5.11$$

So the p-value is smaller than 1%

P-value

$$\text{p-value} = P(|Z| > z_{obs}) = P(|t_{34}| > 5.11)$$

one tail		0.100	0.050	0.025	0.010	0.005
two tails		0.200	0.100	0.050	0.020	0.010
df	31	1.31	1.70	2.04	2.45	2.74
	32	1.31	1.69	2.04	2.45	2.74
	33	1.31	1.69	2.03	2.44	2.73
	34	1.31	1.69	2.03	2.44	2.73
	35	1.31	1.69	2.03	2.44	2.72

$$2.72 < 5.11$$

So the p-value is smaller than 1%

One last example: Disneyland

Disney wants to know if there is significant evidence that their new Paris park grows faster than their CA park.

$$H_0 : \mu_{Paris} \leq \mu_{Cal}$$
$$H_A : \mu_{Paris} > \mu_{Cal}$$

The numbers are not comparable so only the increase in visitors is recorded.

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
CA	NA	-0.2	-1.1	3.8	0.9	-0.8	-0.5	-0.2	0.4	-1.6	0.4	0	0.6	0.96	0.47	0.14	-0.58
Paris	NA	-0.2	-1	1.9	1	0.9	-0.1	0	-0.5	0.2	-1.9	-0.1	0	0	0.4	1.4	0.7

Disneyland `with R`

```
> t.test(disney$cal, disney$paris, paired=T)
```

```
        Paired t-test

data:  disney$cal and disney$paris
t = 0, df = 15, p-value = 1
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.6277834  0.6277834
sample estimates:
mean of the differences
          -1.561251e-17
```

!!!

```
> mean(disney)
```

```
    cal   paris
0.16875 0.16875
```

One last example: Disneyland

Look at what the sum of differences is! Of course we got this number. What we need to look at is the relative change in visitors:

$$\frac{\text{visitors}_{t+1} - \text{visitors}_t}{\text{visitors}_t}$$

where visitors_t is the number of visitors during year t

Year	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
CA	NA	-1.7%	-9.6%	36.9%	6.4%	-5.3%	-3.5%	-1.5%	3.0%	-11.5%	3.3%	0.0%	4.7%	7.2%	3.3%	1.0%	-3.9%
Paris	NA	-2.0%	-10.2%	21.6%	9.3%	7.7%	-0.8%	0.0%	-4.0%	1.7%	-15.6%	-1.0%	0.0%	0.0%	3.9%	13.2%	5.8%

Disneyland `with R`

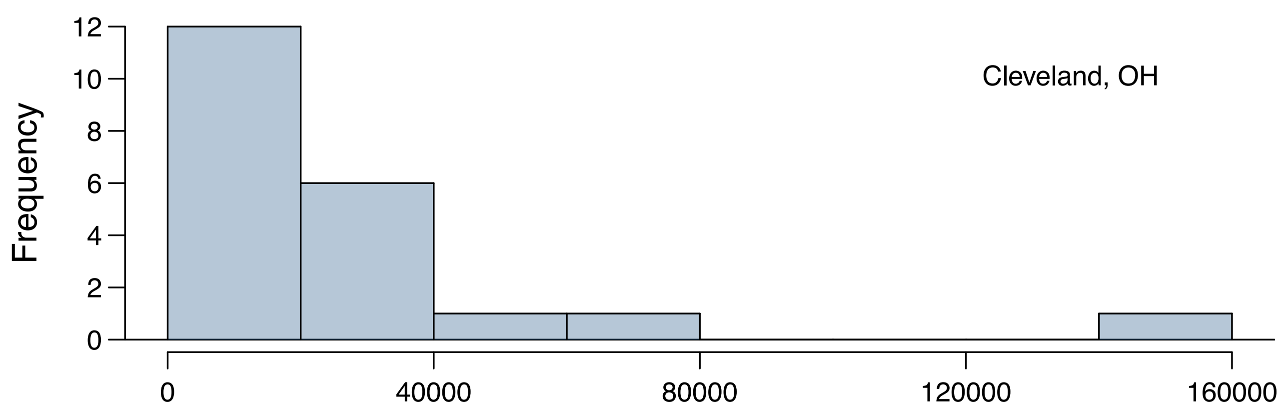
with this new dataset:

```
> t.test(disney$cal, disney$paris, paired=T)
```

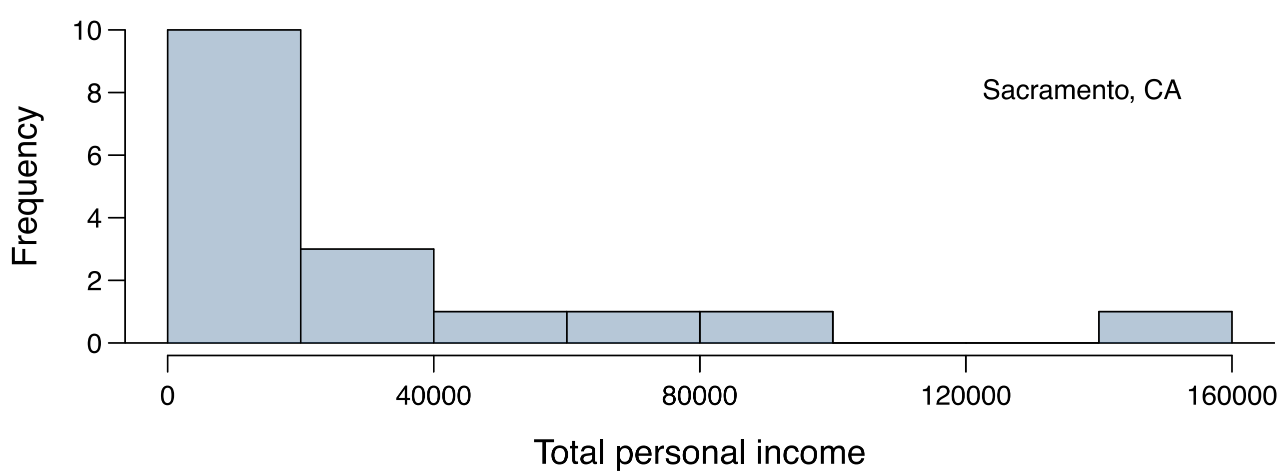
```
        Paired t-test

data:  disney$cal and disney$paris
t = -0.0302, df = 15, p-value = 0.9763
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -0.05069143  0.04927320
sample estimates:
mean of the differences
          -0.0007091129
```

Exercise 6.3.2

Comparing the average total personal income in Cleveland, OH and Sacramento, CA based on a random sample of individuals from the 2000 Census.
Is a t-test appropriate for testing whether or not there is a difference in the average incomes in these two metropolitan cities?



	Cleveland, OH
Mean	$ 26,436
SD	$ 33,239
n	21



	Sacramento, CA
Mean	$ 32,182
SD	$ 40,480
n	17