Tests, Tests, Tests

Eli Gurarie

StatR 101 - Lecture 10b
November 26, 2012

November 26, 2012

Outline

	Distributions	Comparisons	Statistics
1.	Chi-squared(ν)	Proportions	
2.	$\mathcal{T}(\nu)$	Sample means	$\frac{\overline{X}}{s_x/\sqrt{n}}$
3.	$\mathcal{F}(\nu_1, \nu_2)$	Sample variances	$\frac{s_1^2/n_1}{s_2^2/n_2}$

These three distributions are all derived from the normal distribution and are the most widely used null-distributions for hypothesis testing. You will see them pop up frequently in statistical test output.

The one-sample t-statistic

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{\overline{X} - \mu}{s_X / \sqrt{n}}$$

The T-test for comparing means

Question:	Is μ different from μ_0?	Is μ different from μ_0?
Data:	$\overline{X}, n$	$\overline{X}, n, s_x$
Assumptions:	σ_x is known	σ_x is unknown (small sample)
H_0:	$\mu = \mu_0$	
H_A:	$\mu \neq \mu_0$	
Test statistic:	$z_c = \frac{\overline{X} - a}{\sigma_x / \sqrt{n}}$	$t_c = \frac{\overline{X} - \mu_0}{s_x / \sqrt{n}}$
Distribution:	$N(0, 1)$	$T(\nu = n - 1)$
P-value:	$2\,\mathrm{P}(Z > \lvert z_c \rvert)$	$2\,\mathrm{P}(T_\nu > \lvert t_c \rvert)$
α-level	arbitrary!	

This is called the **t-test**. Notice, that it is structurally identical to the z-test, except we use a different distribution. At low degrees of freedom (small n, ν), the tails will be fatter and it will be harder to get significant results.

Example: Dogs of a different size

- Dogs come in different sizes.
- There is an average dog size μ.
- There is some standard deviation of dog size ($\sigma =$? cm).

Example I: What is the average length of a dog - at 95% confidence?

Data: You have 4 (randomly selected) dogs: 89, 91, 112, 124 cm.

$$\overline{X} = 104; s_X = 17; s_{\bar{x}} = 17/\sqrt{4} = 8.5$$

To construct a confidence interval:

$$\widehat{\mu} = \overline{X} \pm t_{c,\nu}\sigma_{\bar{x}}$$

Find the critical value of $t_{c,\nu}$:

- In this case: $\nu = n - 1 = 3$ and C is 95%, but we need the two-tailed value, so we look up $t_{.025,3}$:
- In R: `qt(0.025, df=3) = -3.18` (note, much larger than 1.96)
- So: $\widehat{\mu} = 104 \pm 3.18 * 8.5 = 104 \pm 27$ or: 95% C.I. = (77, 131)

Example I: What is the average length of a dog - at 95% confidence?

Data: You have 4 (randomly selected) dogs: 89, 91, 112, 124 cm.

$\overline{X} = 104; s_X = 17; s_{\bar{x}} = 17/\sqrt{4} = 8.5$

To construct a confidence interval:

$$\widehat{\mu} = \overline{X} \pm t_{c,\nu}\sigma_{\bar{x}}$$

Find the critical value of $t_{c,\nu}$:

- In this case: $\nu = n - 1 = 3$ and C is 95%, but we need the two-tailed value, so we look up $t_{.025,3}$:
- In R: `qt(0.025, df=3) = -3.18` (note, much larger than 1.96)
- So: $\widehat{\mu} = 104 \pm 3.18 * 8.5 = 104 \pm 27$ or: 95% C.I. = (77, 131)

Example I: What is the average length of a dog - at 95% confidence?
Data: You have 4 (randomly selected) dogs: 89, 91, 112, 124 cm.

$\overline{X} = 104; s_X = 17; s_{\bar{x}} = 17/\sqrt{4} = 8.5$

To construct a confidence interval:

$$\widehat{\mu} = \overline{X} \pm t_{c,\nu}\sigma_{\bar{x}}$$

Find the critical value of $t_{c,\nu}$:

- In this case: $\nu = n - 1 = 3$ and C is 95%, but we need the two-tailed value, so we look up $t_{.025,3}$:
- In R: `qt(0.025, df=3) = -3.18` (note, much larger than 1.96)
- So: $\widehat{\mu} = 104 \pm 3.18 * 8.5 = 104 \pm 27$ or: 95% C.I. = (77, 131)

Example I: What is the average length of a dog - at 95% confidence?

Data: You have 4 (randomly selected) dogs: 89, 91, 112, 124 cm.

$\overline{X} = 104; s_X = 17; s_{\bar{x}} = 17/\sqrt{4} = 8.5$

To construct a confidence interval:

$$\widehat{\mu} = \overline{X} \pm t_{c,\nu}\sigma_{\bar{x}}$$

Find the critical value of $t_{c,\nu}$:

- In this case: $\nu = n - 1 = 3$ and C is 95%, but we need the two-tailed value, so we look up $t_{.025,3}$:
- In R: `qt(0.025, df=3) = -3.18` (note, much larger than 1.96)
- So: $\widehat{\mu} = 104 \pm 3.18 * 8.5 = 104 \pm 27$ or: 95% C.I. = (77, 131)

Example II: Hypothesis testing with a single mean

- We know that the global population of domestic dogs has mean length $\mu = 100$ cm.
- We measured length of 16 Sri Lankan strays - and found: $\overline{X} = 92$ cm, and $s_x = 19$.

Question: Are Sri Lankan stray dogs smaller than the average domestic dog (at 5% significance level)?

Example II: Hypothesis testing

1. Null hypothesis: $H_0 : \mu_{stray} = 100$
2. Alt. hypothesis: $H_A : \mu_{stray} < 100$
3. Test statistic:

$$t = \frac{\overline{X} - \mu}{s_{\overline{X}}} = \frac{92 - 100}{19/\sqrt{16}} = \frac{-8}{4.75} = -1.684$$

4. Distribution of t:
   $t \sim$ Students $\mathrm{T}(\nu = 15)$
5. Compare the t statistic to the $t_{.05,15}$ critical value.

Example II: Hypothesis testing

1. Null hypothesis: $H_0 : \mu_{stray} = 100$
2. Alt. hypothesis: $H_A : \mu_{stray} < 100$
3. Test statistic:

$$\begin{aligned} t &= \frac{\overline{X} - \mu}{s_{\overline{X}}} \\ &= \frac{92 - 100}{19/\sqrt{16}} = \frac{-8}{4.75} = -1.684 \end{aligned}$$

4. Distribution of t:
   $t \sim$ Students $\mathrm{T}(\nu = 15)$
5. Compare the t statistic to the $t_{.05,15}$ critical value.

Example II: Hypothesis testing

1. Null hypothesis: $H_0 : \mu_{stray} = 100$
2. Alt. hypothesis: $H_A : \mu_{stray} < 100$
3. Test statistic:

$$
\begin{aligned}
t &= \frac{\overline{X} - \mu}{s_{\overline{X}}} \\
&= \frac{92 - 100}{19/\sqrt{16}} = \frac{-8}{4.75} = -1.684
\end{aligned}
$$

4. Distribution of t:
   $t \sim$ Students $\mathrm{T}(\nu = 15)$
5. Compare the t statistic to the $t_{.05,15}$ critical value.

Looking up $t_{.05,15}$

Recall that this is a one-sided test!

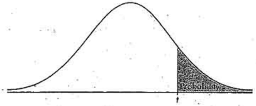


TABLE B: *t*-DISTRIBUTION CRITICAL VALUES

	Tail probability *p*											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	.700	.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	.697	.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	.695	.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	.694	.870	1.079	[illegible]	[illegible]	[illegible]	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	[illegible]	[illegible]	[illegible]	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	[illegible]	**1.753**	[illegible]	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	[illegible]	[illegible]	[illegible]	2.235	2.583	2.921	3.252	3.686	4.015
17	.689	.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	.688	.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.611	3.922
19	.688	.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	.687	.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291
	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%
	Confidence level *C*											

$|t| = 1.684 < 1.753$
$Pr(T_{15} > |t|) > 0.05$

So we **fail to reject** null hypothesis - not enough evidence to state that Sri Lankan strays are truly smaller than the average dog.

Using R: `qt(0.05, df=15)`

Example III: Comparing two small samples

- Say we have 4 thoroughbred dogs: $\overline{X} = 104, s_x = 17$,
- and 16 Sri Lankan stray dogs: $\overline{X} = 92, s_x = 8$,

Question: Is there a difference in their sizes?

Test statistic for 2-sample mean test

- We are interested in the **difference**:

$$D = \overline{X_1} - \overline{X_2}$$

- Basic form of the test statistic is the same: $t = \frac{D - \mu_D}{s_D}$
- ... but there are a few more terms!:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - \mu_D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Note: almost always, the null hypothesis is $\mu_D = 0$ ... this is because we are almost always just interested in comparing the means.
- Note also: for small samples, this statistic behaves properly if X_1 and X_2 have a roughly normal distribution - small sizes mean we can't automatically invoke the central limit theorem.

Test statistic for 2-sample mean test

- We are interested in the **difference**:

$$D = \overline{X_1} - \overline{X_2}$$

- Basic form of the test statistic is the same: $t = \frac{D - \mu_D}{s_D}$
- ... but there are a few more terms!:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - \mu_D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Note: almost always, the null hypothesis is $\mu_D = 0$ ... this is because we are almost always just interested in comparing the means.
- Note also: for small samples, this statistic behaves properly if X_1 and X_2 have a roughly normal distribution - small sizes mean we can't automatically invoke the central limit theorem.

Test statistic for 2-sample mean test

- We are interested in the **difference**:

$$D = \overline{X_1} - \overline{X_2}$$

- Basic form of the test statistic is the same: $t = \frac{D - \mu_D}{s_D}$
- ... but there are a few more terms!:

$$t = \frac{(\overline{X_1} - \overline{X_2}) - \mu_D}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

- Note: almost always, the null hypothesis is $\mu_D = 0$ ... this is because we are almost always just interested in comparing the means.
- Note also: for small samples, this statistic behaves properly if X_1 and X_2 have a roughly normal distribution - small sizes mean we can't automatically invoke the central limit theorem.

What are the *df* for this test?

- There's no exact answer!
- The conservative approach is to use the SMALLER of the two sample sizes (minus 1)
- The best actual approximation is:

$$df = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}\right)^2}{\left(\frac{1}{n_1 - 1}\right)\left(\frac{s_1^2}{n_1}\right)^2 + \left(\frac{1}{n_2 - 1}\right)\left(\frac{s_2^2}{n_2}\right)^2}$$

This is what we call ...

This is what we call ...

Hand Waving!

The two sample t test statistic

If the random samples are drawn, one of size n_1, unknown mean μ_1 and unknown s.d. σ_1, the other of size n_2 with μ_2 and σ_2, also unknown, then to test the hypothesis $H_0 : \mu_1 = \mu_2$, compute the **two sample t statistic**:

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$$

and use P-values or critical values of the t_k distribution, where k is whatever's smaller: $n_1 - 1$ or $n_2 - 1$ (or is approximated by software).

- Note: the term $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ is called: **the pooled standard deviation**.
- Note also: Using the smaller degrees of freedom is the more conservative approach.
- Note finally: The t-distribution is a (very good) approximation of the true distribution of the two-sample test statistic, but it is not *exact* like in the one-sample case.

Back to the dogs

- 4 domestic dogs: $\overline{X} = 104, s_x = 17$
- 16 Sri Lankan strays: $\overline{X} = 92, s_x = 8$,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{104 - 92}{\sqrt{17^2/4 + 8^2/16}} = \frac{12}{8.73} = 1.37$$

Degrees of freedom? $df = (4-1) = 3 < (16-1) = 15$.

Back to the dogs

- 4 domestic dogs: $\overline{X} = 104, s_x = 17$
- 16 Sri Lankan strays: $\overline{X} = 92, s_x = 8$,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{104 - 92}{\sqrt{17^2/4 + 8^2/16}} = \frac{12}{8.73} = 1.37$$

Degrees of freedom? $df = (4-1) = 3 < (16-1) = 15.$

Back to the dogs

- 4 domestic dogs: $\overline{X} = 104, s_x = 17$
- 16 Sri Lankan strays: $\overline{X} = 92, s_x = 8$,

$$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}} = \frac{104 - 92}{\sqrt{17^2/4 + 8^2/16}} = \frac{12}{8.73} = 1.37$$

Degrees of freedom? $df = (4-1) = 3 < (16-1) = 15.$

Recall the question: "Is there a difference in their sizes?"

- So: $H_0 : \mu_1 = \mu_2$, $H_A : \mu_1 \neq \mu_2$
- Test statistic: $t = 1.37$
- Test distribution: $t \sim T(\nu = 3)$
- Significance: $\alpha = 5\%$

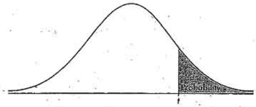


TABLE B: t-DISTRIBUTION CRITICAL VALUES

	Tail probability p											
df	.25	.20	.15	.10	.05	.025	.02	.01	.005	.0025	.001	.0005
1	1.000	1.376	1.963	3.078	6.314	12.71	15.89	31.82	63.66	127.3	318.3	636.6
2	.816	1.061	1.386	1.886	2.92	[illegible]	.849	6.965	9.925	14.09	22.33	31.60
3	.765	.978	1.250	1.638	2.35	**3.182**	.482	4.541	5.841	7.453	10.21	12.92
4	.741	.941	1.190	1.533	2.13	[illegible]	.999	3.747	4.604	5.598	7.173	8.610
5	.727	.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	.718	.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	.711	.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	.706	.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	.703	.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781

$1.37 < 3.182$, so we **fail to reject** H_0.

The two sample T-test for comparing means

Question:	Is μ different from μ_0?	Is μ_1 different from μ_2?
Test:	Single sample t-test	Two sample t-test
Data:	$\overline{X}, n, s$	$\overline{X_1}, n_1, s_1, \overline{X_2}, n_2, s_2$
Assumptions:	Roughly normal distributions of X, small sample	
H_0:	$\mu = \mu_0$	$\mu_1 = \mu_2$
H_A:	$\mu \neq \mu_0$	$\mu_1 \neq \mu_2$
Test statistic:	$t = \frac{\overline{X} - \mu_0}{s_X / \sqrt{n}}$	$t = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$
Distribution:	$T(\nu = n - 1)$	$T(\nu \approx min(n_1 - 1, n_2 - 1))$
P-value:	$2\,\mathrm{P}(T_\nu > \lvert t \rvert)$	
α-level:	arbitrary!	

Example IV: Taxi fleet

- Louie De Palma has a big fleet of taxis, and is very cheap.
- To save some money, he wants to see if Gasoline A is more efficient than Gasoline B (at 95% confidence).
- He reads a statistics book, and randomly assigns 50 cars to Gasoline A, and 50 cars to Gasoline B.
-

Data:	Sample size	Mean mileage	SD
A	50	25	5.00
B	50	26	4.00

- Quick test shows:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{-1}{\sqrt{\frac{25}{50} + \frac{16}{50}}} = -1.10$$

$$|t| = 1.10 < 2.01; \quad \Pr(T_{49} > |t|) > 0.05$$

- Obviously, there is no difference between the gasolines!

Example IV: Taxi fleet

- Louie De Palma has a big fleet of taxis, and is very cheap.
- To save some money, he wants to see if Gasoline A is more efficient than Gasoline B (at 95% confidence).
- He reads a statistics book, and randomly assigns 50 cars to Gasoline A, and 50 cars to Gasoline B.
-

Data:	Sample size	Mean mileage	SD
A	50	25	5.00
B	50	26	4.00

- Quick test shows:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{-1}{\sqrt{\frac{25}{50} + \frac{16}{50}}} = -1.10$$

$$|t| = 1.10 < 2.01; \quad \Pr(T_{49} > |t|) > 0.05$$

- Obviously, there is no difference between the gasolines!

Example IV: Taxi fleet

- Louie De Palma has a big fleet of taxis, and is very cheap.
- To save some money, he wants to see if Gasoline A is more efficient than Gasoline B (at 95% confidence).
- He reads a statistics book, and randomly assigns 50 cars to Gasoline A, and 50 cars to Gasoline B.
-

Data:	Sample size	Mean mileage	SD
A	50	25	5.00
B	50	26	4.00

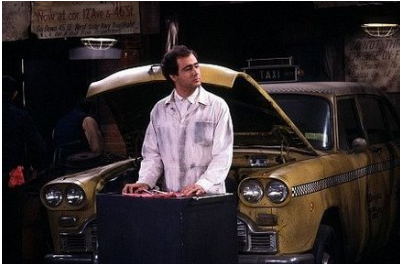
- Quick test shows:

$$t = \frac{\overline{x_1} - \overline{x_2}}{\sqrt{s_1^2/n_1 + s_2^2/n_2}} = \frac{-1}{\sqrt{\frac{25}{50} + \frac{16}{50}}} = -1.10$$

$$|t| = 1.10 < 2.01; \quad \Pr(T_{49} > |t|) > 0.05$$

- Obviously, there is no difference between the gasolines!

But Wait!



- Latka Gravas (who is very smart) says: *"But wait! Gas B looked a little better than Gas A - but the standard deviation was very wide."*
    - Why? Because taxis (and taxi drivers) are all very different.
- Maybe a better experiment is to use the same cab but assign Gas A and Gas B to the same cab on different days!

A new experiment

Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
4	2.22	26.32	-24.10
5	30.11	29.56	0.55
6	5.55	26.60	-21.05
7	22.23	22.93	-0.70
8	19.78	20.23	-0.45
9	33.45	33.95	-0.50
10	25.22	26.01	-0.79
$\overline{X}$	**25.2**	**25.8**	**-0.60**
s_x	**4.27**	**4.10**	**0.61**

- Note: to do this right - you randomize the order of Gas A and Gas B (by flipping a coin) - but control for driver.
- Note: the standard deviations and means are similar as before, but the *difference* has a very small standard deviation
- Note: The sample size here is quite a bit smaller than before.

A new experiment

Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
4	25.22	26.32	-1.10
5	30.11	29.56	0.55
6	25.55	26.60	-1.05
7	22.23	22.93	-0.70
8	19.78	20.23	-0.45
9	33.45	33.95	-0.50
10	25.22	26.01	-0.79
$\overline{X}$	**25.2**	**25.8**	**-0.60**
s_x	**4.27**	**4.10**	**0.61**

- The differences d_i is a *single measure of the difference of each taxi*.
- We can narrow the question to: "Is the true $\mu_d = 0$?" and apply a single sample t-test!
- Let's calculate a 95% CI:

$$\begin{aligned} \mu_d &= \overline{d} \pm t_{c,\nu} \frac{s_d}{\sqrt{n}} \\ &= -.6 \pm (2.26)(\frac{.61}{\sqrt{10}}) \\ &= -.6 \pm 0.44 \end{aligned}$$

- So: $-1.04 \leq \mu_d \leq -0.16$ with 95% confidence - and there is good evidence that Gas B really is better than Gas A!

A new experiment

Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
4	25.22	26.32	-1.10
5	30.11	29.56	0.55
6	25.55	26.60	-1.05
7	22.23	22.93	-0.70
8	19.78	20.23	-0.45
9	33.45	33.95	-0.50
10	25.22	26.01	-0.79
$\overline{X}$	**25.2**	**25.8**	**-0.60**
s_x	**4.27**	**4.10**	**0.61**

- The differences d_i is a *single measure of the difference of each taxi*.
- We can narrow the question to: "Is the true $\mu_d = 0$?" and apply a single sample t-test!
- Let's calculate a 95% CI:

$$\begin{aligned}\mu_d &= \overline{d} \pm t_{c,\nu}\frac{s_d}{\sqrt{n}} \\ &= -.6 \pm (2.26)(\frac{.61}{\sqrt{10}}) \\ &= -.6 \pm 0.44\end{aligned}$$

- So: $-1.04 \leq \mu_d \leq -0.16$ with 95% confidence - and there is good evidence that Gas B really is better than Gas A!

A new experiment

Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
4	25.22	26.32	-1.10
5	30.11	29.56	0.55
6	25.55	26.60	-1.05
7	22.23	22.93	-0.70
8	19.78	20.23	-0.45
9	33.45	33.95	-0.50
10	25.22	26.01	-0.79
$\overline{X}$	**25.2**	**25.8**	**-0.60**
s_x	**4.27**	**4.10**	**0.61**

- The differences d_i is a *single measure of the difference of each taxi*.
- We can narrow the question to: "Is the true $\mu_d = 0$?" and apply a single sample t-test!
- Let's calculate a 95% CI:

$$\begin{aligned} \mu_d &= \overline{d} \pm t_{c,\nu} \frac{s_d}{\sqrt{n}} \\ &= -.6 \pm (2.26)(\frac{.61}{\sqrt{10}}) \\ &= -.6 \pm 0.44 \end{aligned}$$

- So: $-1.04 \leq \mu_d \leq -0.16$ with 95% confidence - and there is good evidence that Gas B really is better than Gas A!

A new experiment

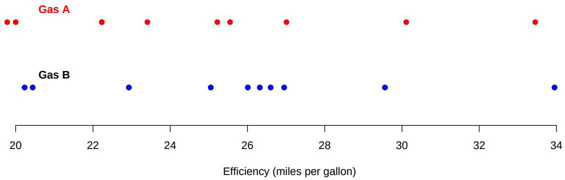
Cab	Gas A	Gas B	Difference
1	27.01	26.95	0.06
2	20.00	20.44	-0.44
3	23.41	25.05	-1.64
4	25.22	26.32	-1.10
5	30.11	29.56	0.55
6	25.55	26.60	-1.05
7	22.23	22.93	-0.70
8	19.78	20.23	-0.45
9	33.45	33.95	-0.50
10	25.22	26.01	-0.79
$\overline{X}$	**25.2**	**25.8**	**-0.60**
s_x	**4.27**	**4.10**	**0.61**

- The differences d_i is a *single measure of the difference of each taxi*.
- We can narrow the question to: "Is the true $\mu_d = 0$?" and apply a single sample *t*-test!
- Let's calculate a 95% CI:

$$\begin{aligned} \mu_d &= \overline{d} \pm t_{c,\nu} \frac{s_d}{\sqrt{n}} \\ &= -.6 \pm (2.26)(\frac{.61}{\sqrt{10}}) \\ &= -.6 \pm 0.44 \end{aligned}$$

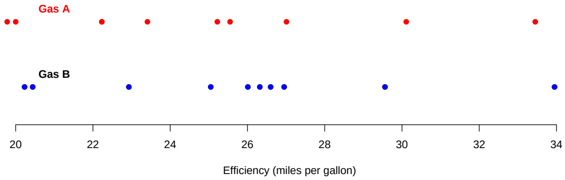
- So: $-1.04 \leq \mu_d \leq -0.16$ with 95% confidence - and there is good evidence that Gas B really is better than Gas A!

Visualizing paired Comparisons

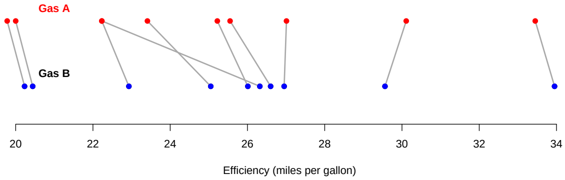


Variability swamps the signal

Visualizing paired Comparisons



Variability swamps the signal



Paired comparison of means

Question:	Is μ different from μ_0?	Is μ_d different from 0?
Test:	Single sample t-test	Paired comparison test
Data:	$\overline{X}, n, s_x$	$\overline{d}, n, s_d$
Assumptions:	Roughly normal distributions of X, small sample	
H_0:	$\mu = \mu_0$	$\mu_d = 0$
H_A:	$\mu \neq \mu_0$	$\mu_d \neq 0$
Test statistic:	$t = \frac{\overline{X} - \mu_0}{s_x / \sqrt{n}}$	$t = \frac{\overline{d}}{s_d / \sqrt{n}}$
Distribution:	$T(\nu = n - 1)$	
P-value:	$2\,\mathrm{P}(T_\nu > \lvert t \rvert)$	
α-level:	arbitrary!	

Paired (or "matched pair") Comparison

- Is one of the most effective ways to compare treatments while controlling for natural variability.
- Allows for much greater power, because: $s_d \ll s_{x_1}$.
- Used for natural pairings where the effect might be smaller than the variability, for example:
  - Effectiveness of two hand creams: compare Right and Left hands (randomized) to control for different skin types.
  - Aggressive behavior of dementia patients on full moon: compare Full moon and non-full moon days to control for variability in aggression.
  - Effect of caffeine (or flower smelling) on student's brains: compare treatment and lack of treatment with a randomized repeated measure.

Paired (or "matched pair") Comparison

- Is one of the most effective ways to compare treatments while controlling for natural variability.
- Allows for much greater power, because: $s_d \ll s_{x_1}$.
- Used for natural pairings where the effect might be smaller than the variability, for example:
  - Effectiveness of two hand creams: compare Right and Left hands (randomized) to control for different skin types.
  - Aggressive behavior of dementia patients on full moon: compare Full moon and non-full moon days to control for variability in aggression.
  - Effect of caffeine (or flower smelling) on student's brains: compare treatment and lack of treatment with a randomized repeated measure.

Paired (or "matched pair") Comparison

- Is one of the most effective ways to compare treatments while controlling for natural variability.
- Allows for much greater power, because: $s_d \ll s_{x_1}$.
- Used for natural pairings where the effect might be smaller than the variability, for example:
    - Effectiveness of two hand creams: compare Right and Left hands (randomized) to control for different skin types.
    - Aggressive behavior of dementia patients on full moon: compare Full moon and non-full moon days to control for variability in aggression.
    - Effect of caffeine (or flower smelling) on student's brains: compare treatment and lack of treatment with a randomized repeated measure.

Paired (or "matched pair") Comparison

- Is one of the most effective ways to compare treatments while controlling for natural variability.
- Allows for much greater power, because: $s_d \ll s_{x_1}$.
- Used for natural pairings where the effect might be smaller than the variability, for example:
  - Effectiveness of two hand creams: compare Right and Left hands (randomized) to control for different skin types.
  - Aggressive behavior of dementia patients on full moon: compare Full moon and non-full moon days to control for variability in aggression.
  - Effect of caffeine (or flower smelling) on student's brains: compare treatment and lack of treatment with a randomized repeated measure.

Paired (or "matched pair") Comparison

- Is one of the most effective ways to compare treatments while controlling for natural variability.
- Allows for much greater power, because: $s_d \ll s_{x_1}$.
- Used for natural pairings where the effect might be smaller than the variability, for example:
  - Effectiveness of two hand creams: compare Right and Left hands (randomized) to control for different skin types.
  - Aggressive behavior of dementia patients on full moon: compare Full moon and non-full moon days to control for variability in aggression.
  - Effect of caffeine (or flower smelling) on student's brains: compare treatment and lack of treatment with a randomized repeated measure.

Paired (or "matched pair") Comparison

- Is one of the most effective ways to compare treatments while controlling for natural variability.
- Allows for much greater power, because: $s_d \ll s_{x_1}$.
- Used for natural pairings where the effect might be smaller than the variability, for example:
    - Effectiveness of two hand creams: compare Right and Left hands (randomized) to control for different skin types.
    - Aggressive behavior of dementia patients on full moon: compare Full moon and non-full moon days to control for variability in aggression.
    - Effect of caffeine (or flower smelling) on student's brains: compare treatment and lack of treatment with a randomized repeated measure.

The F-statistic for comparing sample variances

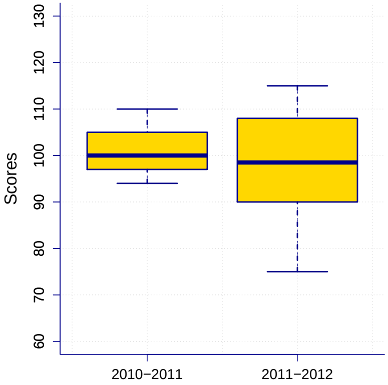
$$\begin{aligned} F_{obs} &= \frac{s_1^2}{s_2^2} \\ &\sim \mathcal{F}(\nu_1 = n_1 - 1, \nu_2 = n_2 - 1) \end{aligned}$$

The F-statistic allows us to compare two *sample variances*.

Example 5: NBA games

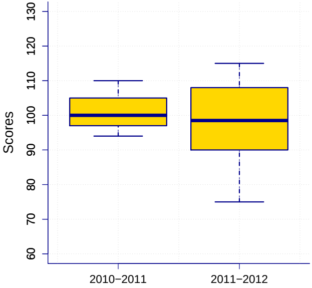
Question: Are teams playing less consistently this season than last season because of a compressed schedule?

Game	**2010-2011**	**2011-2012**
1	100	111
2	95	108
3	97	99
4	101	94
5	100	115
6	94	100
7	110	88
8	105	75
9	98	98
10	109	90
means	100.90	97.80
s.d.	6.0	12.0



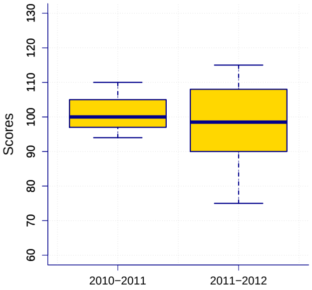
An example: NBA games

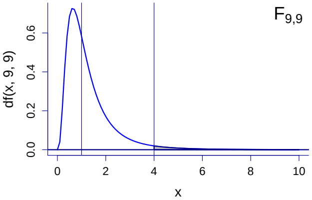
- Hypotheses
  - H_0: $\sigma_1 = \sigma_2$;
  - H_1: $\sigma_2 > \sigma_1$;
- Data:
  - $s_1 = 6, n = 10$
  - $s_2 = 12, n = 10$



An example: NBA games

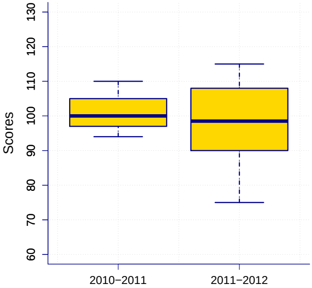
- Hypotheses
  - H_0: $\sigma_1 = \sigma_2$;
  - H_1: $\sigma_2 > \sigma_1$;
- Data:
  - $s_1 = 6, n = 10$
  - $s_2 = 12, n = 10$
- Test statistic:
  - $f_{obs} = s_2^2 / s_1^2 = 4$
  - $f_{obs} \sim F_{n_1 - 1 = 9, n_2 - 1 = 9}$
- P-value:
  - $\Pr(F_{9,9} > f_{obs}) = 0.025$

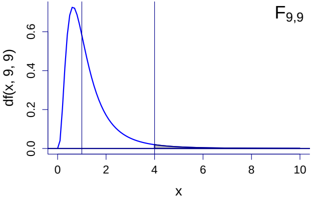




An example: NBA games

- Hypotheses
  - H_0: $\sigma_1 = \sigma_2$;
  - H_1: $\sigma_2 > \sigma_1$;
- Data:
  - $s_1 = 6, n = 10$
  - $s_2 = 12, n = 10$
- Test statistic:
  - $f_{obs} = s_2^2 / s_1^2 = 4$
  - $f_{obs} \sim F_{n_1 - 1 = 9, n_2 - 1 = 9}$
- P-value:
  - $\Pr(F_{9,9} > f_{obs}) = 0.025$





Conclusion: reject null-hypothesis, games ARE more inconsistent this year than last

Testing the differences between variances

Question:	Is μ_1 greater than μ_2?	Is σ_1 greater than σ_2?
Test:	Two sample t-test	F-test
Data:	$\overline{X_1}, n_1, s_1, \overline{X_2}, n_2, s_2$	s_1, s_2
Assumptions:	X is roughly normal, samples are small	
H_0:	$\mu_1 = \mu_2$	$\sigma_1 = \sigma_2$
H_A:	$\mu_1 > \mu_2$	$\sigma_1 > \sigma_2$
Test statistic:	$t_{test} = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}$	$F_{test} = \frac{s_1^2}{s_2^2}$
Distribution:	$T(\nu \approx min(n_1 - 1, n_2 - 1))$	$F(\nu_1 = n_1 - 1, \nu_2 = n_2 - 2$
P-value:	$\mathrm{P}(T_\nu > \lvert t_{test} \rvert)$	$\mathrm{P}(F_{\nu_1,\nu_2} > F_{test})$

Robustness

- All statistical test rest on assumptions.
  - Most common assumption: the test statistic has a normal distribution
  - variances are equal in populations being compared
  - samples are drawn independently and randomly
- What if the assumptions are violated - do the tests still work?
  - i.e. Can they provide *valid inference* from a sample?
- If yes, the the test is **robust** to violations of the assumptions
  - A test may be robust to some violations, but not others.
  - Violations include: presence of outliers, inappropriate distributions, unequal variances, etc.

Robustness

- All statistical test rest on assumptions.
    - Most common assumption: the test statistic has a normal distribution
    - variances are equal in populations being compared
    - samples are drawn independently and randomly
- What if the assumptions are violated - do the tests still work?
    - i.e. Can they provide *valid inference* from a sample?
- If yes, the the test is **robust** to violations of the assumptions
    - A test may be robust to some violations, but not others.
    - Violations include: presence of outliers, inappropriate distributions, unequal variances, etc.

Robustness

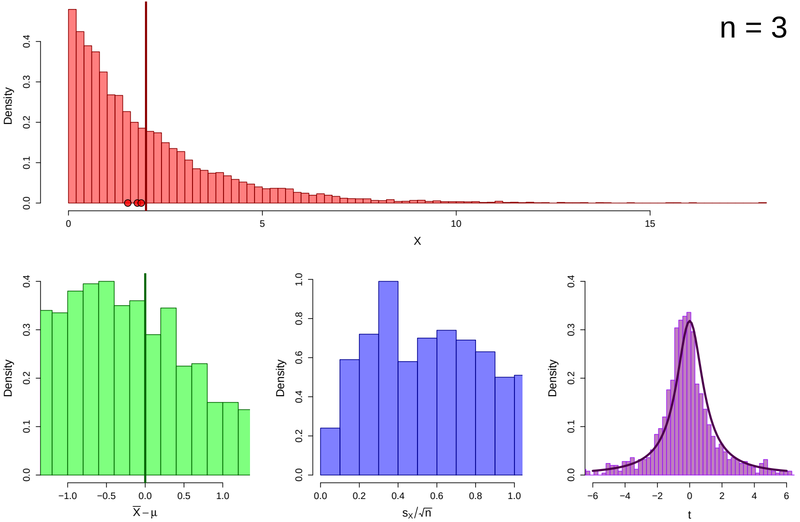
- All statistical test rest on assumptions.
    - Most common assumption: the test statistic has a normal distribution
    - variances are equal in populations being compared
    - samples are drawn independently and randomly
- What if the assumptions are violated - do the tests still work?
    - i.e. Can they provide *valid inference* from a sample?
- If yes, the the test is **robust** to violations of the assumptions
    - A test may be robust to some violations, but not others.
    - Violations include: presence of outliers, inappropriate distributions, unequal variances, etc.

Robustness

Some rules of thumb:

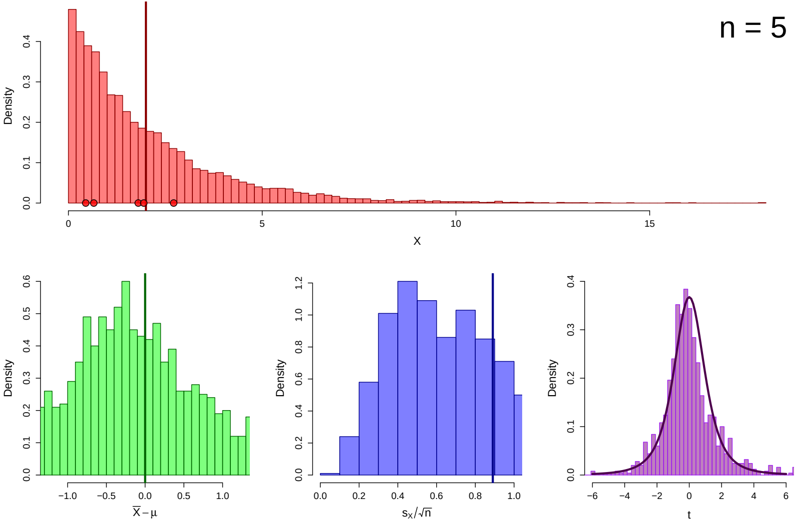
- Two-sample T-procedures are more robust than one-sample T-procedures.
- T-tests are most robust when both sample sizes are equal and both sample distributions are similar.
- ... but even when we deviate from this, two-sample tests tend to remain quite robust.
- F-tests tend to be very **sensitive** (opposite of **robust**) to non-normality assumptions.

T robust!

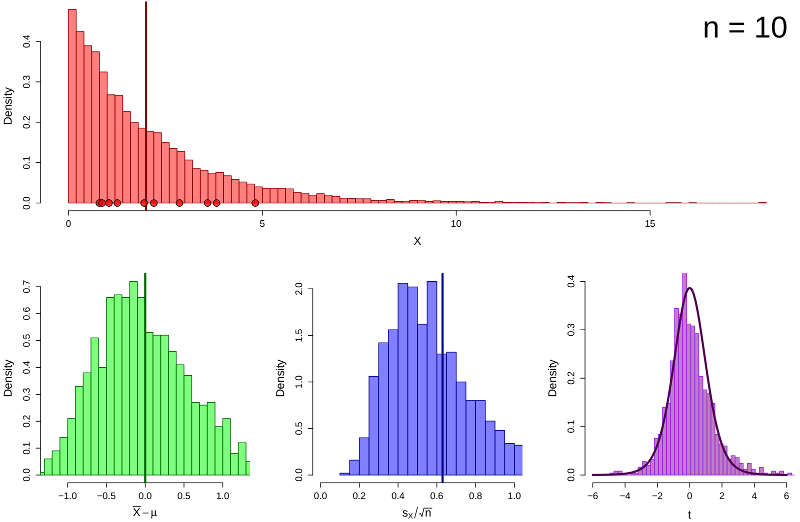


n = 3
Density
X
0
5
10
15
0.0
0.1
0.2
0.3
0.4
$\overline{X} - \mu$
−1.0
−0.5
0.0
0.5
1.0
$s_X/\sqrt{n}$
0.2
0.4
0.6
0.8
1.0
t
−6
−4
−2
0
2
4
6

T robust!

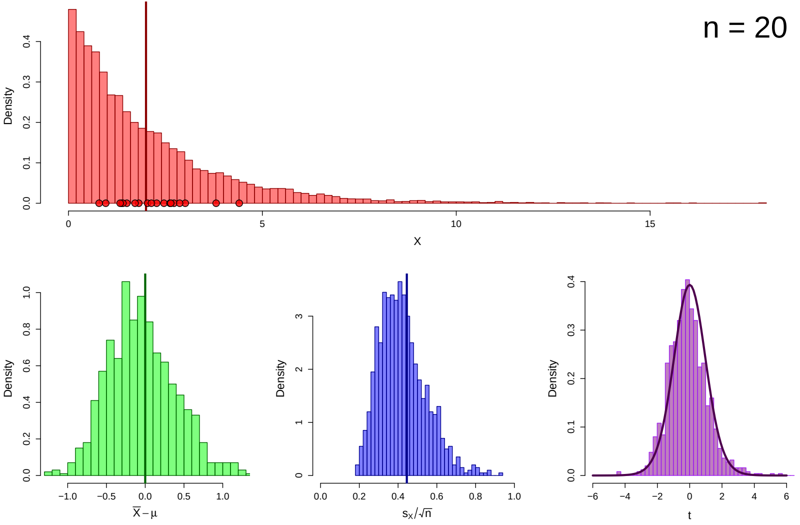

n = 5
Density
X
0
5
10
15
0.0
0.1
0.2
0.3
0.4
$\overline{X} - \mu$
−1.0
−0.5
0.0
0.5
1.0
0.6
0.5
$s_X/\sqrt{n}$
0.2
0.4
0.6
0.8
1.2
1.0
t
−6
−4
−2
0
2
4
6

T robust!

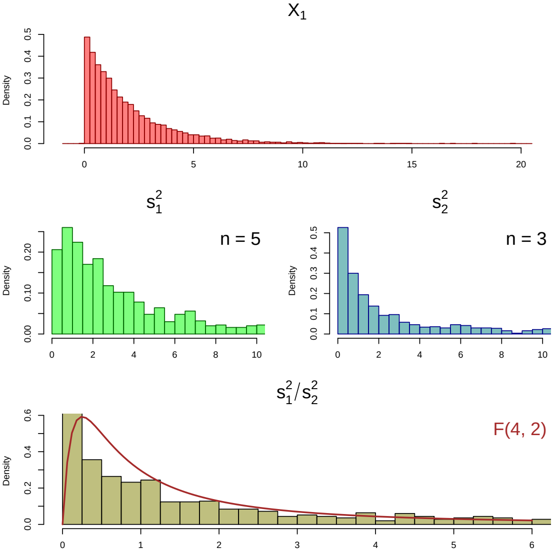


n = 10
Density
X
0
5
10
15
0.0
0.1
0.2
0.3
0.4
$\overline{X} - \mu$
−1.0
−0.5
0.0
0.5
1.0
0.0
0.1
0.2
0.3
0.4
0.5
0.6
0.7
$s_X/\sqrt{n}$
0.0
0.2
0.4
0.6
0.8
1.0
0.5
1.5
2.0
t
−6
−4
−2
0
2
4
6

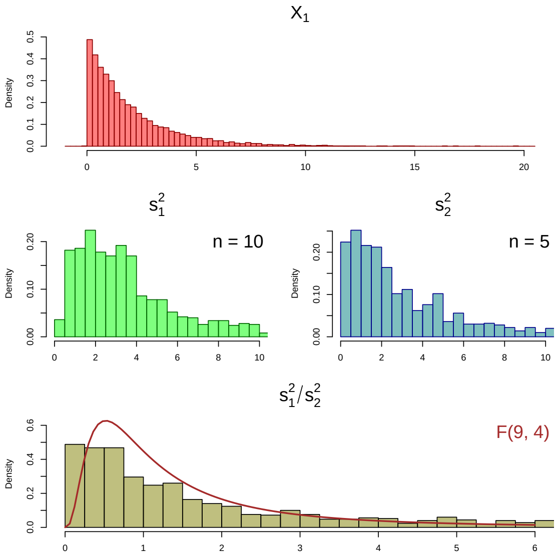
T robust!



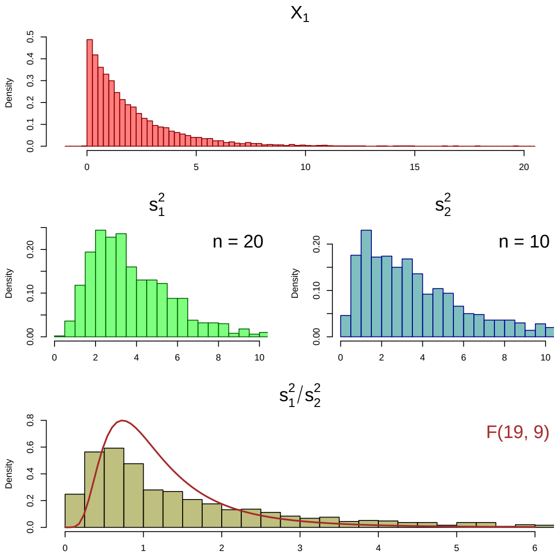
F robust?



F robust?

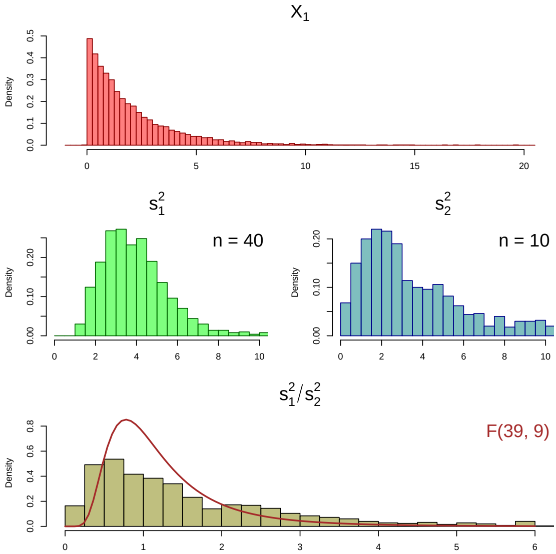


F robust?



X_1
Density
0.0
0.1
0.2
0.3
0.4
0.5
0
5
10
15
20
s_1^2
n = 20
s_2^2
n = 10
0.00
0.10
0.20
0
2
4
6
8
10
s_1^2/s_2^2
F(19, 9)
0.0
0.2
0.4
0.6
0.8
0
1
2
3
4
5
6

F robust?



X_1
Density
0.0 0.1 0.2 0.3 0.4 0.5
0 5 10 15 20
s_1^2
s_2^2
n = 40
n = 10
Density
0.00 0.10 0.20
0 2 4 6 8 10
s_1^2/s_2^2
F(39, 9)
Density
0.0 0.2 0.4 0.6 0.8
0 1 2 3 4 5 6