Math 361

Comparing two population means – Inv. 4.2

Last Time

The **Card Shuffle Randomization** was used estimate the p-value for a **simulation** test of

$$H_0: \mu_1 = \mu_2 \quad \text{vs.} \quad H_a: \mu_1 > \mu_2$$

*Card Shuffling was used to mimic the random assignment of volunteers to sleep deprivation or not in the **experiment.***

*We also noted that the simulated null distribution looked normal and so, if we use the sample SD's instead of the population SD's, the distribution should be **approximately t**.*

Inv. 4.2 – NBA Salaries

Question: Do Western and Eastern Conference NBA players make the same salary, on average?

Let's try a simulation test of the null hypothesis of no difference in mean salaries.

Is the data collection like a **card shuffle**, that is, should I picture someone being randomly assigned to the East or West coast?

Study Design and Simulation Tests

Randomized Experimental study – randomly assigned participants to one or the other EV group

Simulate data assuming H_0 is true = card shuffle

Observational study – no random assignment to one or the other EV group, just observation of naturally occurring groups.

Simulate random samples from the population assuming H_0 is true

Inv. 4.2: NBA Salaries by Conference

Descriptive statistics: compare numerical summaries

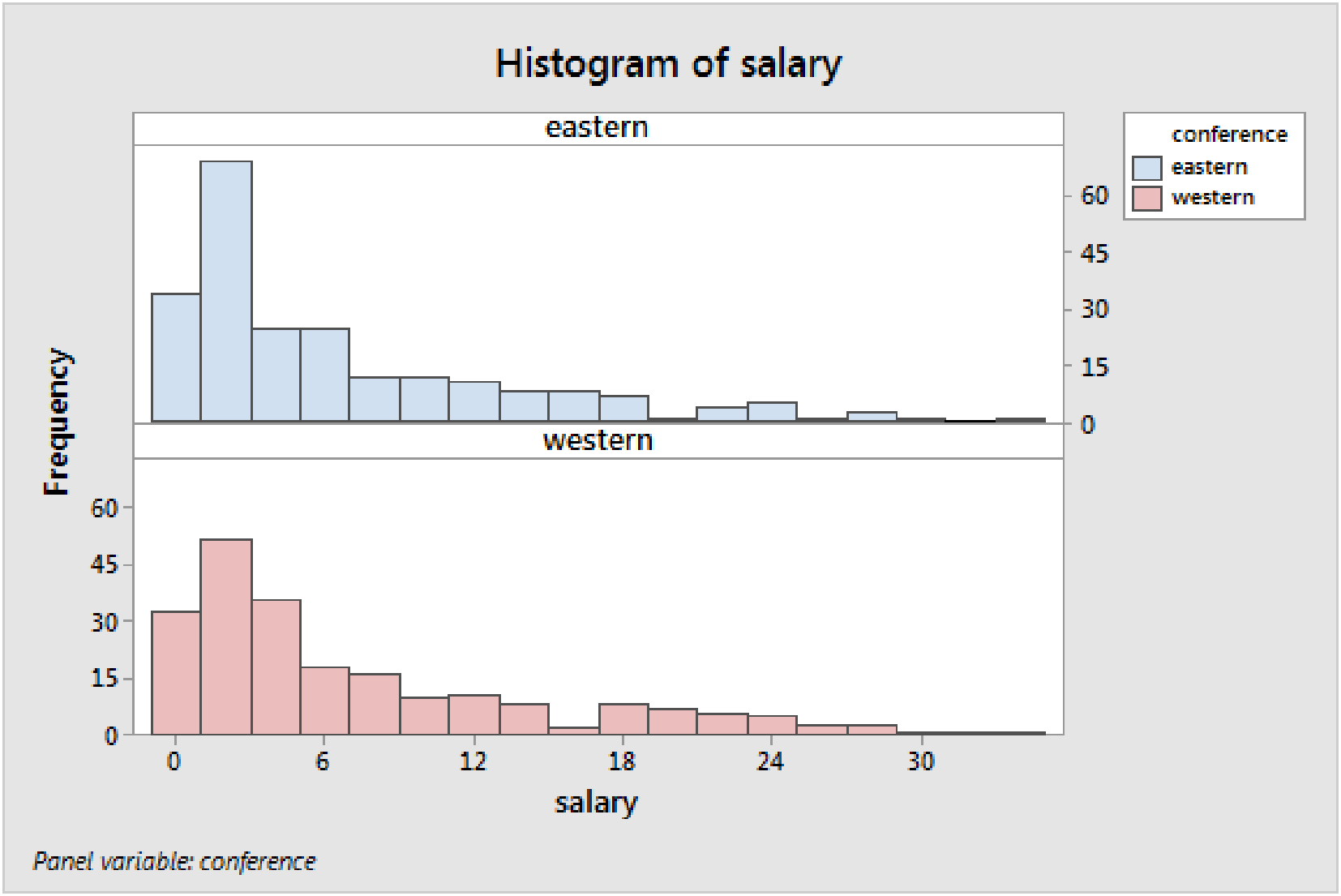
Statistics

Variable	conference	N	Mean	StDev	Variance	Median	IQR
salary	eastern	227	6.614	7.006	49.081	3.807	8.363
	western	221	7.421	7.757	60.164	4.000	9.763

Salaries are in millions of dollars

Inv. 4.2: NBA Salaries by Conference

Descriptive statistics – compare graphs



Inv. 4.2: NBA Salaries by Conference

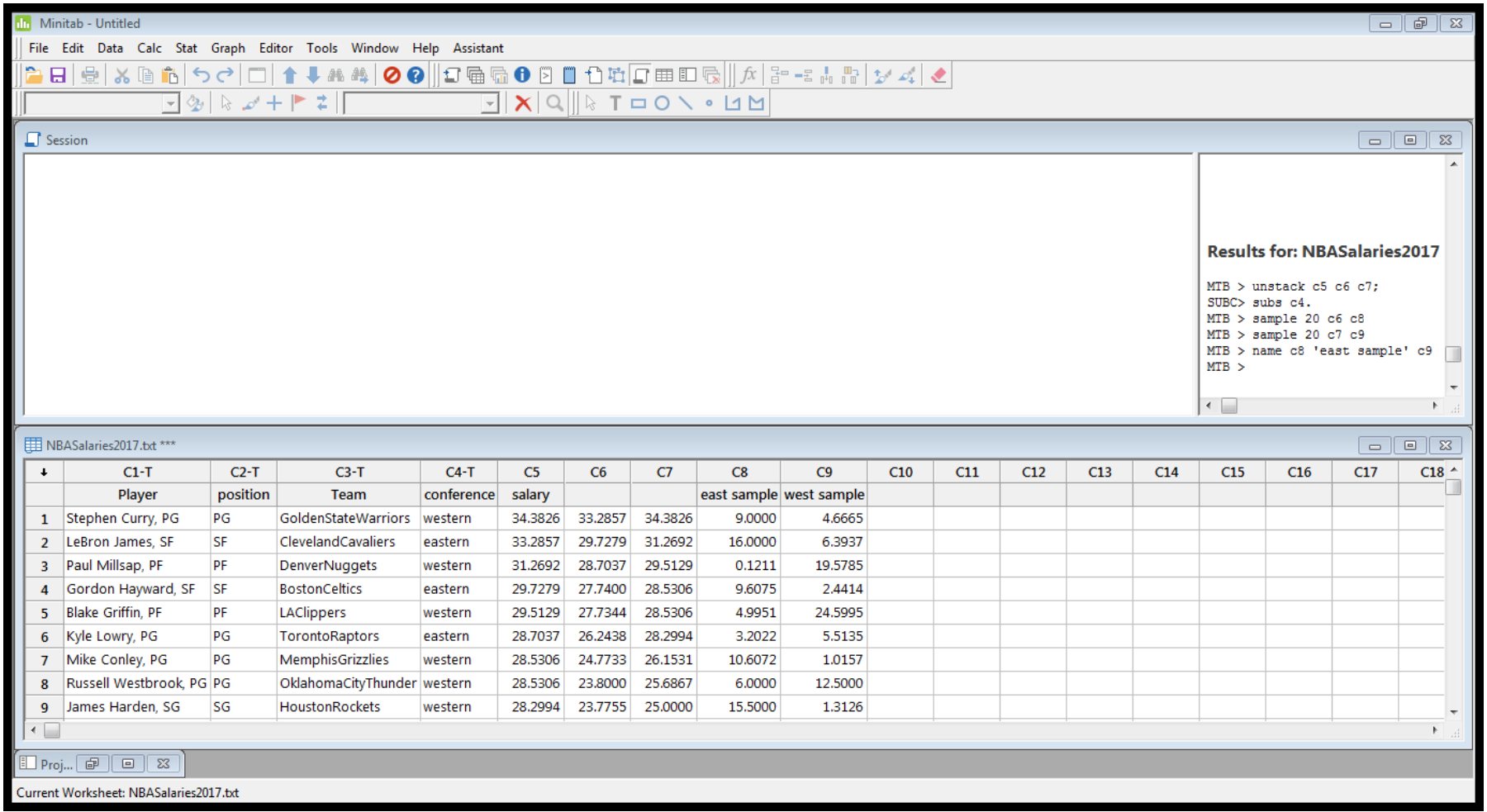
Let's treat the salaries from each conference in 2017-18 as two populations and explore the *distribution of differences in sample means* assuming the null hypothesis is true.

To do so, you'll need to use Minitab, R or Excel

Note: you can use Minitab on any on-campus computer or download a free 30 day trial.

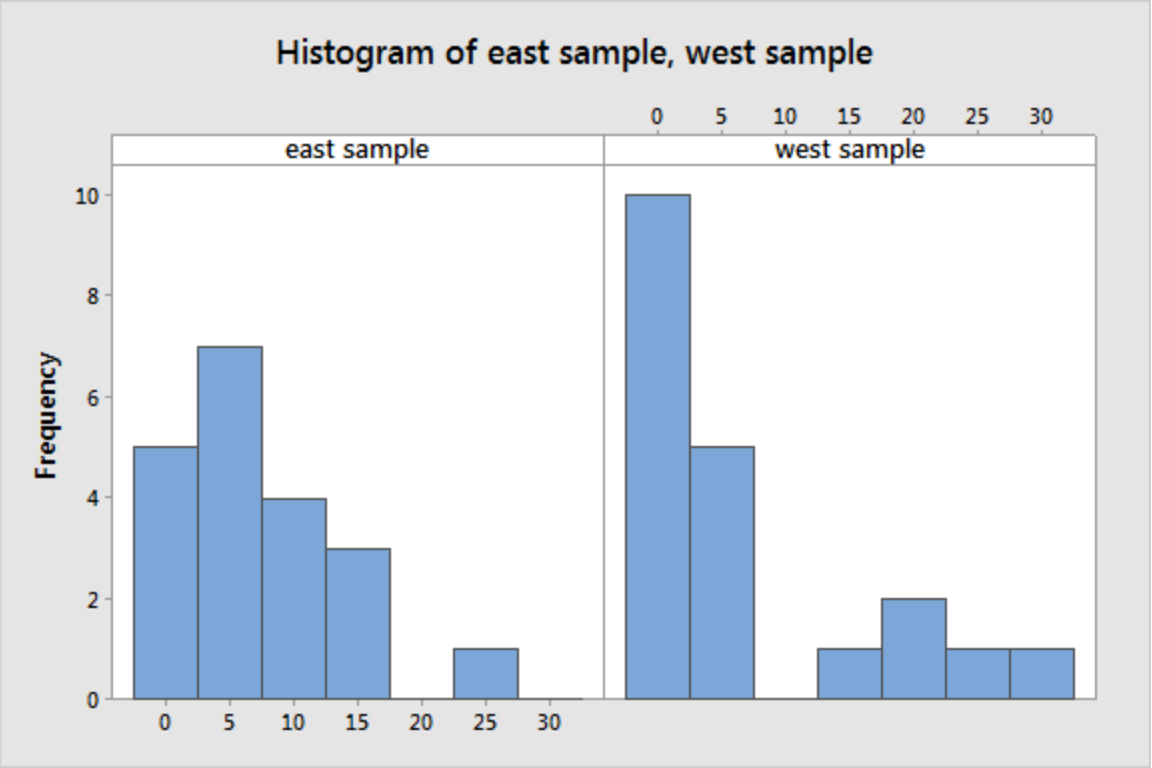
Simulation in Minitab

- Open NBASalaries2017.txt in Minitab
- Follow steps on page 258



	C1-T	C2-T	C3-T	C4-T	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17	C18
	Player	position	Team	conference	salary			east sample	west sample									
1	Stephen Curry, PG	PG	GoldenStateWarriors	western	34.3826	33.2857	34.3826	9.0000	4.6665									
2	LeBron James, SF	SF	ClevelandCavaliers	eastern	33.2857	29.7279	31.2692	16.0000	6.3937									
3	Paul Millsap, PF	PF	DenverNuggets	western	31.2692	28.7037	29.5129	0.1211	19.5785									
4	Gordon Hayward, SF	SF	BostonCeltics	eastern	29.7279	27.7400	28.5306	9.6075	2.4414									
5	Blake Griffin, PF	PF	LAClippers	western	29.5129	27.7344	28.5306	4.9951	24.5995									
6	Kyle Lowry, PG	PG	TorontoRaptors	eastern	28.7037	26.2438	28.2994	3.2022	5.5135									
7	Mike Conley, PG	PG	MemphisGrizzlies	western	28.5306	24.7733	26.1531	10.6072	1.0157									
8	Russell Westbrook, PG	PG	OklahomaCityThunder	western	28.5306	23.8000	25.6867	6.0000	12.5000									
9	James Harden, SG	SG	HoustonRockets	western	28.2994	23.7755	25.0000	15.5000	1.3126									

A pair of samples, both of size 20



Statistics

Variable	N	Mean	StDev	Median	IQR
east sample	20	7.66	6.15	6.33	9.59
west sample	20	7.68	9.28	3.55	10.17

Generate 1000 pairs of samples, compute the difference in sample means for each

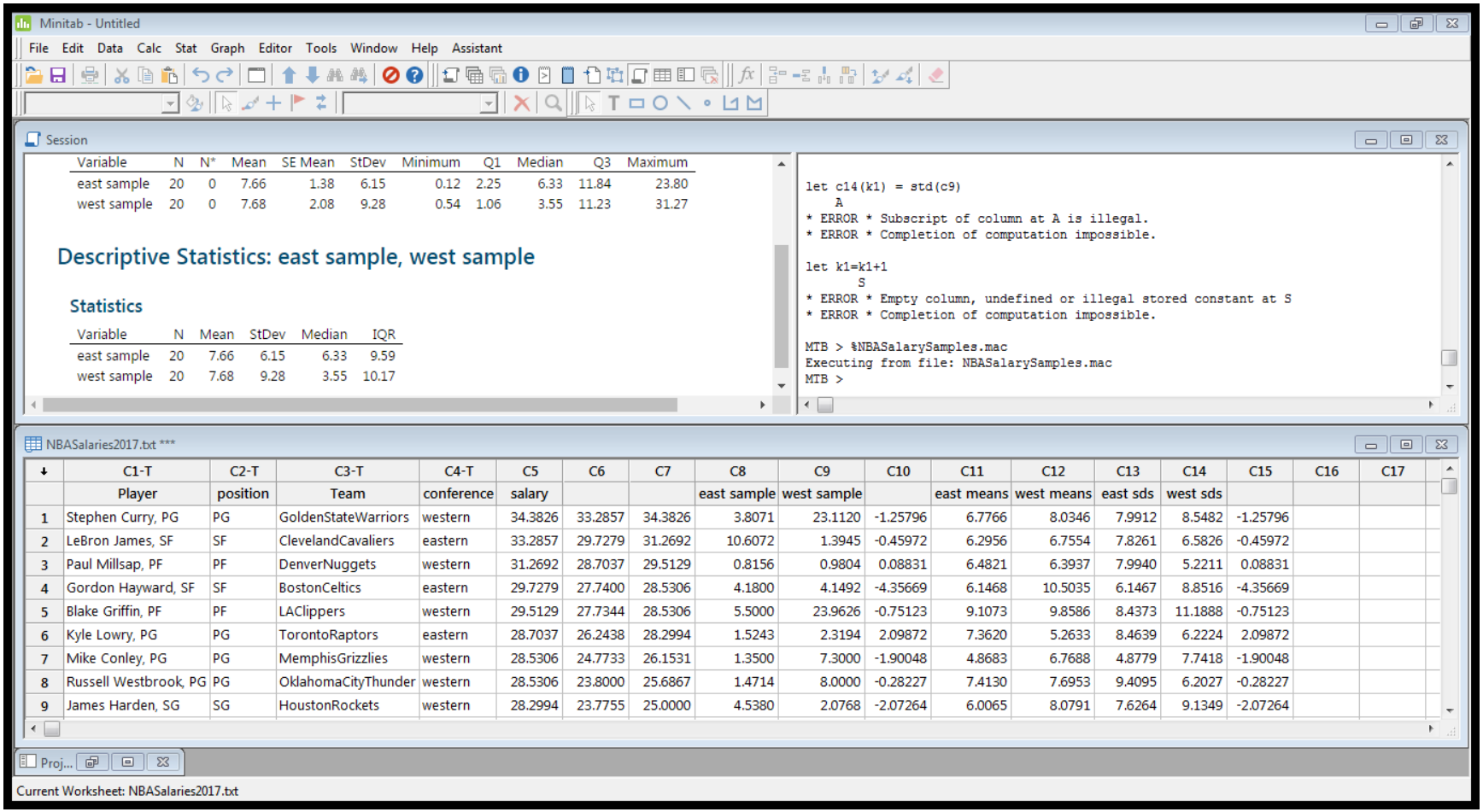
Save NBASalarySamples.MAC in the same location as your copy of NBASalaries2017.txt

Go to **Editor** – **command line** and type

```
MTB > %NBASalarySamples.mac
```

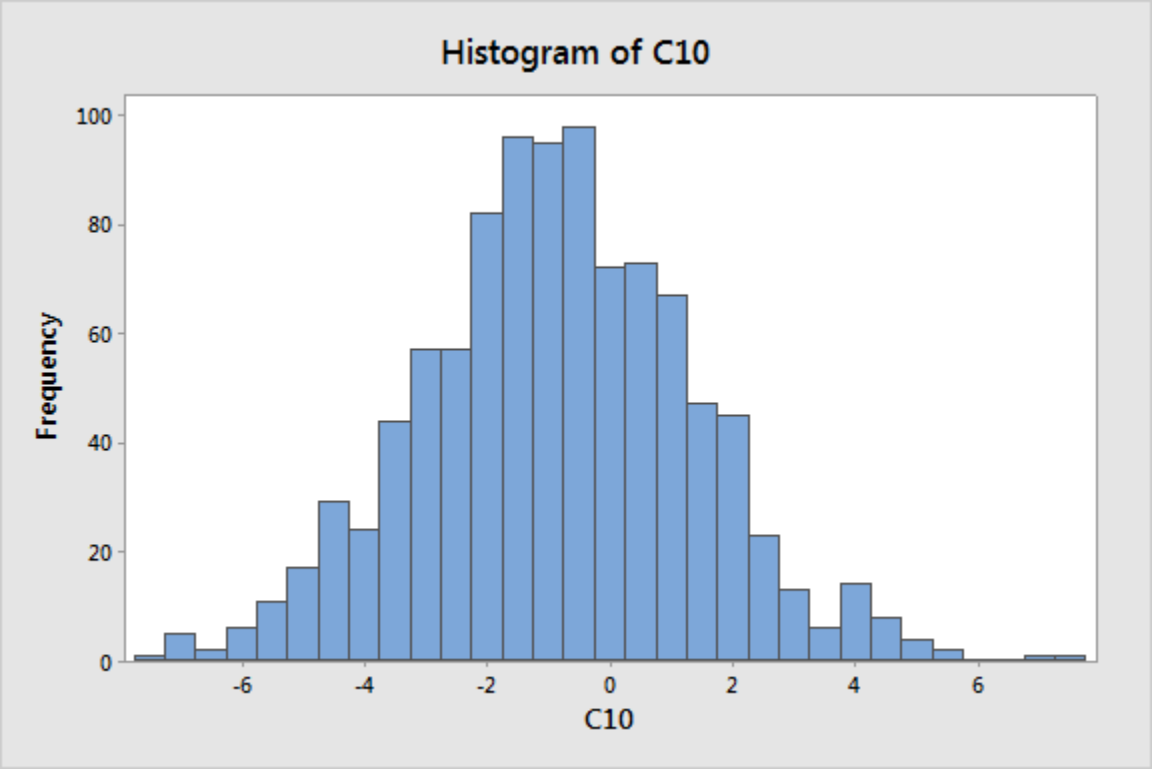
The differences in sample means will be in C10

Result in Minitab



	C1-T	C2-T	C3-T	C4-T	C5	C6	C7	C8	C9	C10	C11	C12	C13	C14	C15	C16	C17
	Player	position	Team	conference	salary			east sample	west sample		east means	west means	east sds	west sds			
1	Stephen Curry, PG	PG	GoldenStateWarriors	western	34.3826	33.2857	34.3826	3.8071	23.1120	-1.25796	6.7766	8.0346	7.9912	8.5482	-1.25796		
2	LeBron James, SF	SF	ClevelandCavaliers	eastern	33.2857	29.7279	31.2692	10.6072	1.3945	-0.45972	6.2956	6.7554	7.8261	6.5826	-0.45972		
3	Paul Millsap, PF	PF	DenverNuggets	western	31.2692	28.7037	29.5129	0.8156	0.9804	0.08831	6.4821	6.3937	7.9940	5.2211	0.08831		
4	Gordon Hayward, SF	SF	BostonCeltics	eastern	29.7279	27.7400	28.5306	4.1800	4.1492	-4.35669	6.1468	10.5035	6.1467	8.8516	-4.35669		
5	Blake Griffin, PF	PF	LAClippers	western	29.5129	27.7344	28.5306	5.5000	23.9626	-0.75123	9.1073	9.8586	8.4373	11.1888	-0.75123		
6	Kyle Lowry, PG	PG	TorontoRaptors	eastern	28.7037	26.2438	28.2994	1.5243	2.3194	2.09872	7.3620	5.2633	8.4639	6.2224	2.09872		
7	Mike Conley, PG	PG	MemphisGrizzlies	western	28.5306	24.7733	26.1531	1.3500	7.3000	-1.90048	4.8683	6.7688	4.8779	7.7418	-1.90048		
8	Russell Westbrook, PG	PG	OklahomaCityThunder	western	28.5306	23.8000	25.6867	1.4714	8.0000	-0.28227	7.4130	7.6953	9.4095	6.2027	-0.28227		
9	James Harden, SG	SG	HoustonRockets	western	28.2994	23.7755	25.0000	4.5380	2.0768	-2.07264	6.0065	8.0791	7.6264	9.1349	-2.07264		

Distribution of differences in sample means



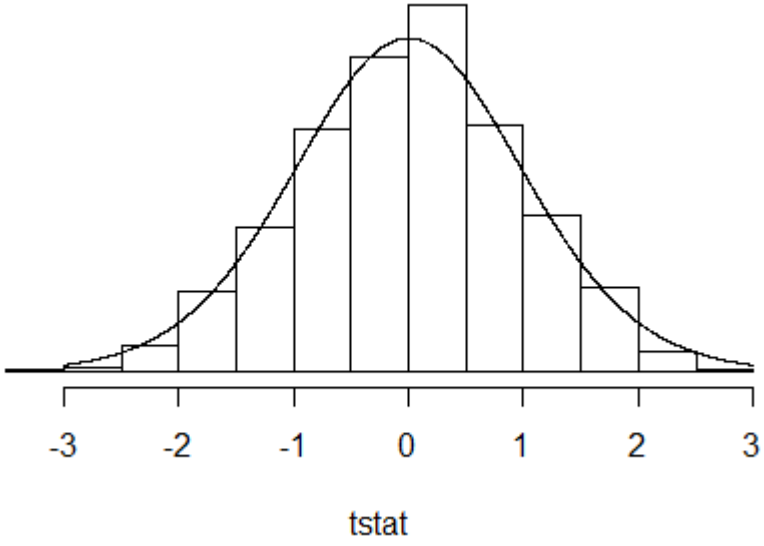
Recap

- Luckily, the distribution of the *differences in sample means* follows a very predictable pattern

CLT:
Mean: μ_1-μ_2
SD: sqrt($\sigma_1^2/n_1 + \sigma_2^2/n_2$)
Approximately normal if populations not too skewed or samples too small

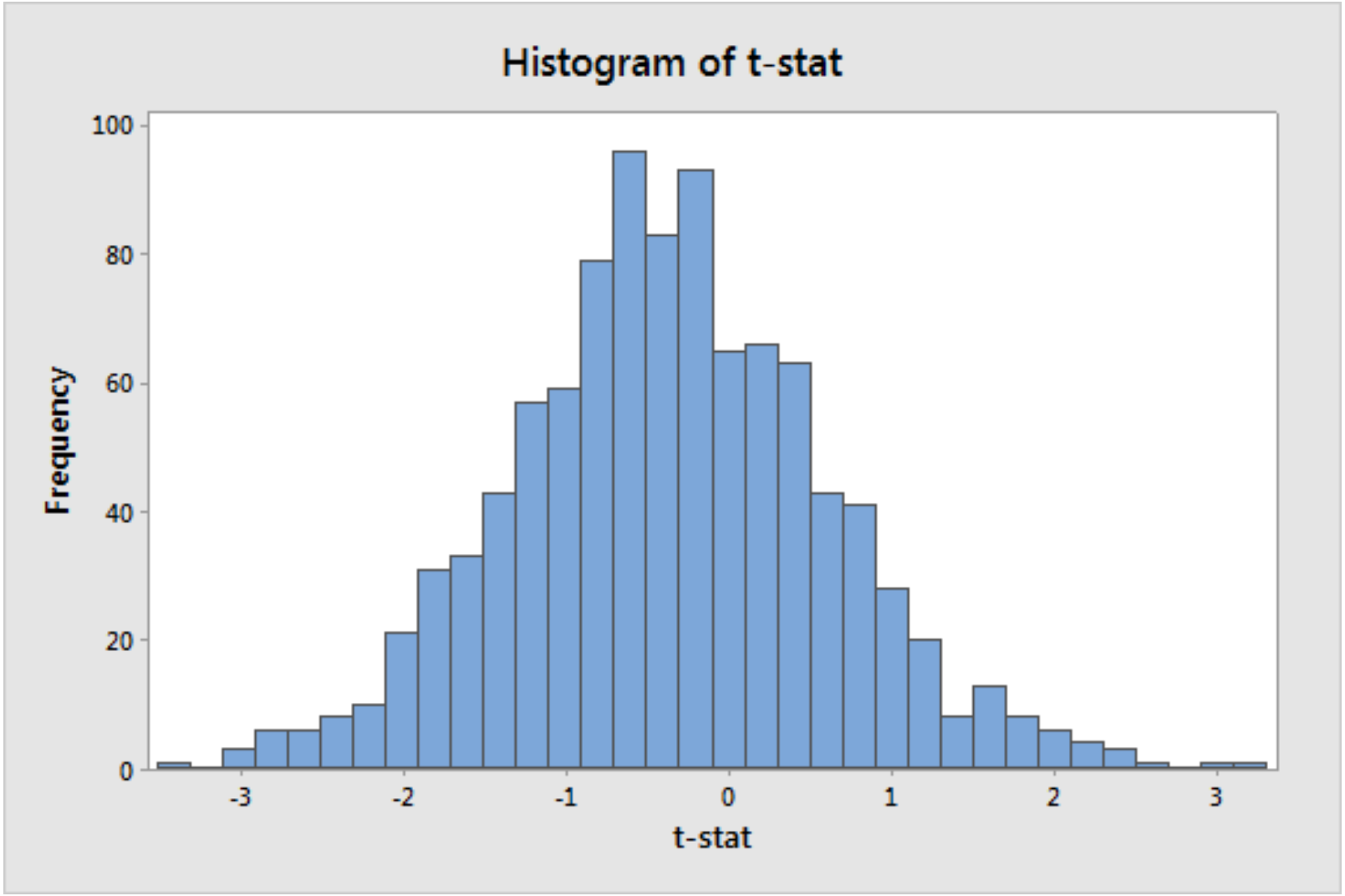
Recap

- Which means, when we use the sample standard deviations to replace the population SDs, the standardized statistic will be well-modelled by a *t*-distribution



The appropriate degrees of freedom are a little complicated, but we'll let the computer deal with that

Inv. 4.2, part j: t-statistics



Summary of Two-sample *t* Procedures

Parameter: $\mu_1 - \mu_2$ = the difference in the population means

To test H_0: $\mu_1 - \mu_2 = \delta_0$

Test statistic: $t_0 = \dfrac{(\bar{x}_1 - \bar{x}_2) - \delta_0}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$

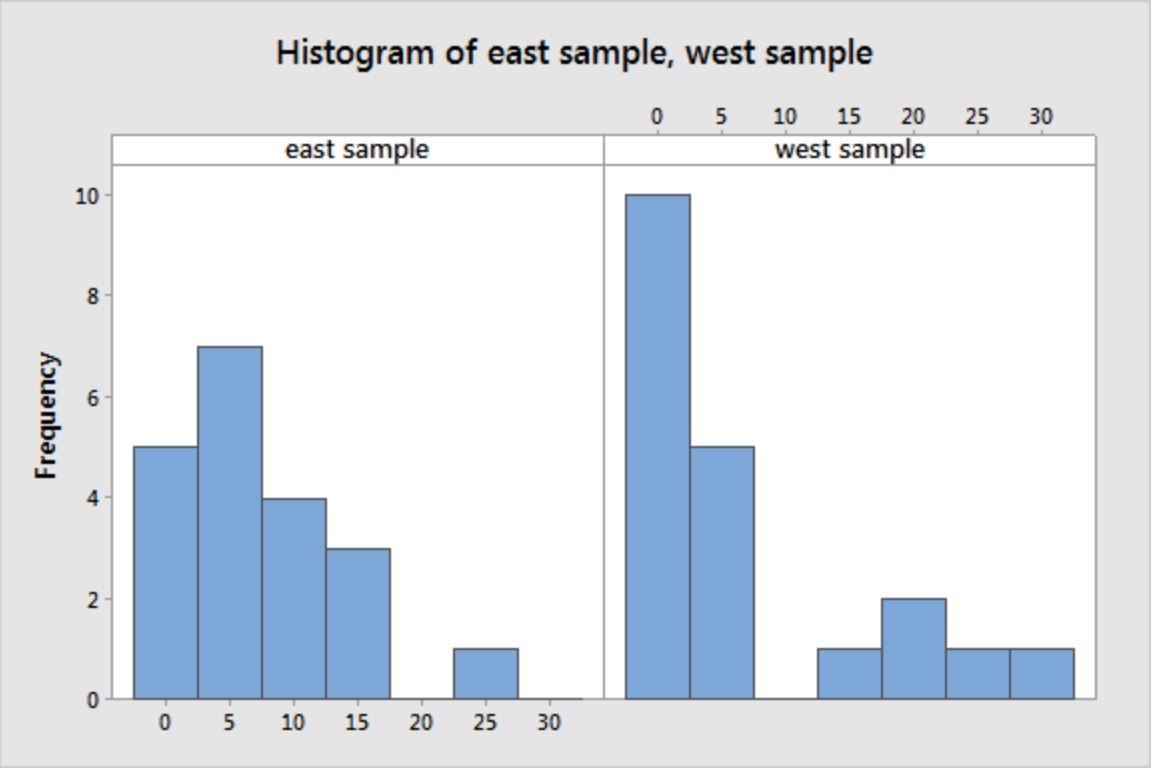
t-**Confidence interval for** $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \times \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

Approximate degrees of freedom: Compare this to a *t*-distribution with degrees of freedom equal to the smaller of the two samples sizes minus one: $min(n_1, n_2) - 1$.

Technical conditions: These procedures are considered valid if the sample distributions are reasonably symmetric or the sample sizes are both at least 20.

A pair of samples, both of size 20



Statistics

Variable	N	Mean	StDev	Median	IQR
east sample	20	7.66	6.15	6.33	9.59
west sample	20	7.68	9.28	3.55	10.17

Using the **TBI Applet** – valid because both samples sizes are greater than or equal to 20

Rossman/Chance Applet Collection

Theory-Based Inference

Scenario: Two means

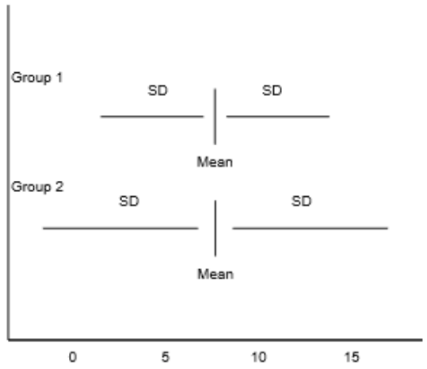
☐ Paste Data

Group 1		Group 2	
n:	20	n:	20
mean, $\bar{x}$:	7.66	mean, $\bar{x}$:	7.68
sample sd, s:	6.15	sample sd, s:	9.28

Calculate

Reset

Sample Data



(Group1 - Group2)

$\bar{x}_1 - \bar{x}_2 = -0.02$

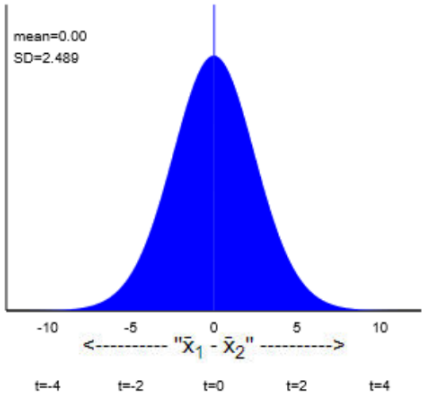
Theory-Based Inference

☑ Test of significance

H_o: $\mu_1 - \mu_2$ = 0

H_a: $\mu_1 - \mu_2$ ≠ 0

Calculate



standardized statistic t = -0.01 df = 32.99

p-value 0.9936

☑ Confidence interval

confidence level 95 % Calculate CI

(-5.0848, 5.0448)

df = 32.99

Interpretations

P-value: There is a 99% chance of seeing two random samples of Eastern and Western players have a mean salary difference of $20,000 or more if there is actually no difference in mean salaries between the two conferences.

T-statistic: The observed mean salary difference of 0.02 (i.e. $20,000) is 0.01 SD's from 0.

95% CI: I am 95% confident that the actual mean salary difference of Eastern and Western players is between -5.08 million and 5.04 million dollars.

Conclusion addressing generalizability and causation

With such a high probability of seeing my data assuming the null hypothesis is true, I conclude there's no evidence against the idea that Western and Eastern conference NBA players have different salaries on average.

Random samples were used so I'm fairly certain my results generalize to all NBA players in 2017-18.

The observational study design does **not** allow for cause and effect relationships to be concluded because of the possibility of confounding variables. For example, better players may earn higher salaries and be more likely to want to live on the West Coast.

Conclusion

With such a high probability of seeing my data assuming the null hypothesis is true, I conclude there's no evidence against the idea that Western and Eastern conference NBA players have different salaries on average.

Notice that I actually do know which hypothesis is true for 2017-18: Western players make an average of 1.8 million dollars more!

Which kind of error did I make, type I or type II?

Conclusion

With such a high probability of seeing my data assuming the null hypothesis is true, I conclude there's no evidence against the idea that Western and Eastern conference NBA players have different salaries on average.

Notice that I actually do know which hypothesis is true for 2017-18: Western players make an average of 1.8 million dollars more!

I falsely accepted the null when alternative was true, type II

Probably this was due to low power: with a small sample size it is unlikely I'll be able to correctly reject the null.

Notice the 95% CI did not lead me astray: 1.8 is between -5.08 and 5.04 million dollars

More Practice: Low carb diet vs. Conventional

A study by Foster el al., reported in *The New England Journal of Medicine* (May, 2003), investigated the effectiveness of a popular "low-carb" diet. The researchers randomly assigned 63 obese men and women to either a low-carbohydrate, high-protein, high-fat (Atkins) diet or a low-calorie, high-carbohydrate, low-fat (conventional) diet. The mean amount of weight lost, as percent of body weight, after 3 months, 6 months and 12 months are shown in the table below. (The baseline weight was carried forward in the case of missing values.)

- **Is this an observational study or an experiment? Explain.**
- **Identify the explanatory and response variables.**
- **Report the relevant hypotheses (in symbols) for testing whether the mean weight losses differ significantly between the two diets.**

RV = weight loss

EV = diet

μ_1 = mean weight loss in diet 1

H_0: $\mu_1 = \mu_2$ → $\mu_1 - \mu_2 = 0$

H_a: $\mu_1 > \mu_2$ → $\mu_1 - \mu_2 > 0$

Time	Diet	Sample size	Mean	SD
3 months	Low-carb	33	6.8	5.0
	Conventional	30	2.7	3.7
6 months	Low-carb	33	7.0	6.5
	Conventional	30	3.2	5.6
12 months	Low-carb	33	4.4	6.7
	Conventional	30	2.5	6.3

More Practice: Low carb diet vs. Conventional

• Is this an observational study or an experiment? Explain.

This is an experiment, because the researchers randomly assigned subjects to either the low-carb diet or the conventional diet.

• Identify the explanatory and response variables.

The explanatory variable is the type of diet (low-carb or traditional) to which the subject was assigned. The response variable is the amount of weight loss as a percentage of body weight.

• Report the relevant hypotheses (in symbols) for testing whether the mean weight losses differ significantly between the two diets.

The hypotheses are H_0: $\mu_{lowcarb} = \mu_{conventional}$ vs. H_a: $\mu_{lowcarb} \neq \mu_{conventional}$.

Interpret the p-values

Time	Diet	Sample size	Mean	SD	P-value
3 months	Low-carb	33	6.8	5.0	0.00051
	Conventional	30	2.7	3.7	
6 months	Low-carb	33	7.0	6.5	0.0187
	Conventional	30	3.2	5.6	
12 months	Low-carb	33	4.4	6.7	0.2556
	Conventional	30	2.5	6.3	

Conclusion addressing generalizability and causation?

Time	Diet	Sample size	Mean	SD	P-value
3 months	Low-carb	33	6.8	5.0	0.00051
	Conventional	30	2.7	3.7	
6 months	Low-carb	33	7.0	6.5	0.0187
	Conventional	30	3.2	5.6	
12 months	Low-carb	33	4.4	6.7	0.2556
	Conventional	30	2.5	6.3	

Conclusion addressing generalizability and causation?

Initially the low carb diet lead to more weight loss on average but by 12 months there was not a significant difference.

People were not randomly chosen to participate so I wouldn't generalize beyond people willing to volunteer for diet studies.

Random assignment was used to assign volunteers to the low carb or conventional diet so the study design does allow for the detection of a cause and effect relationship between diet and weight loss.