Hypothesis testing 1

Independent t-test

Two samples: goal

- We want to test if systolic blood pressure (*sbp*) is significantly different between boys and girls (*sex*)

- You need to determine if independent t-test or Mann-Whitney-U test is appropriate. Which data characteristics do you need to check?

In the *t*-test comparing the means of two independent samples, the following assumptions should be met:

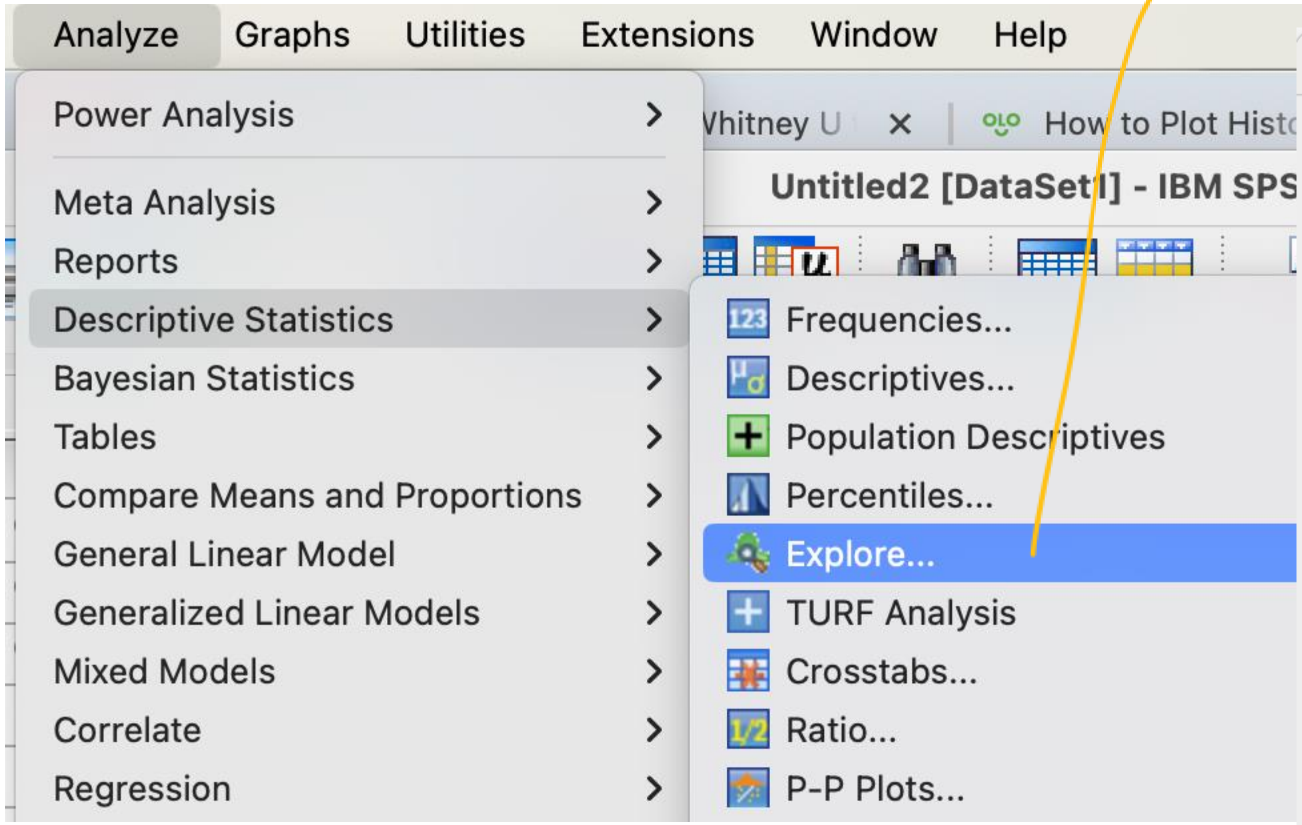
NOT the means of the two samples!!!

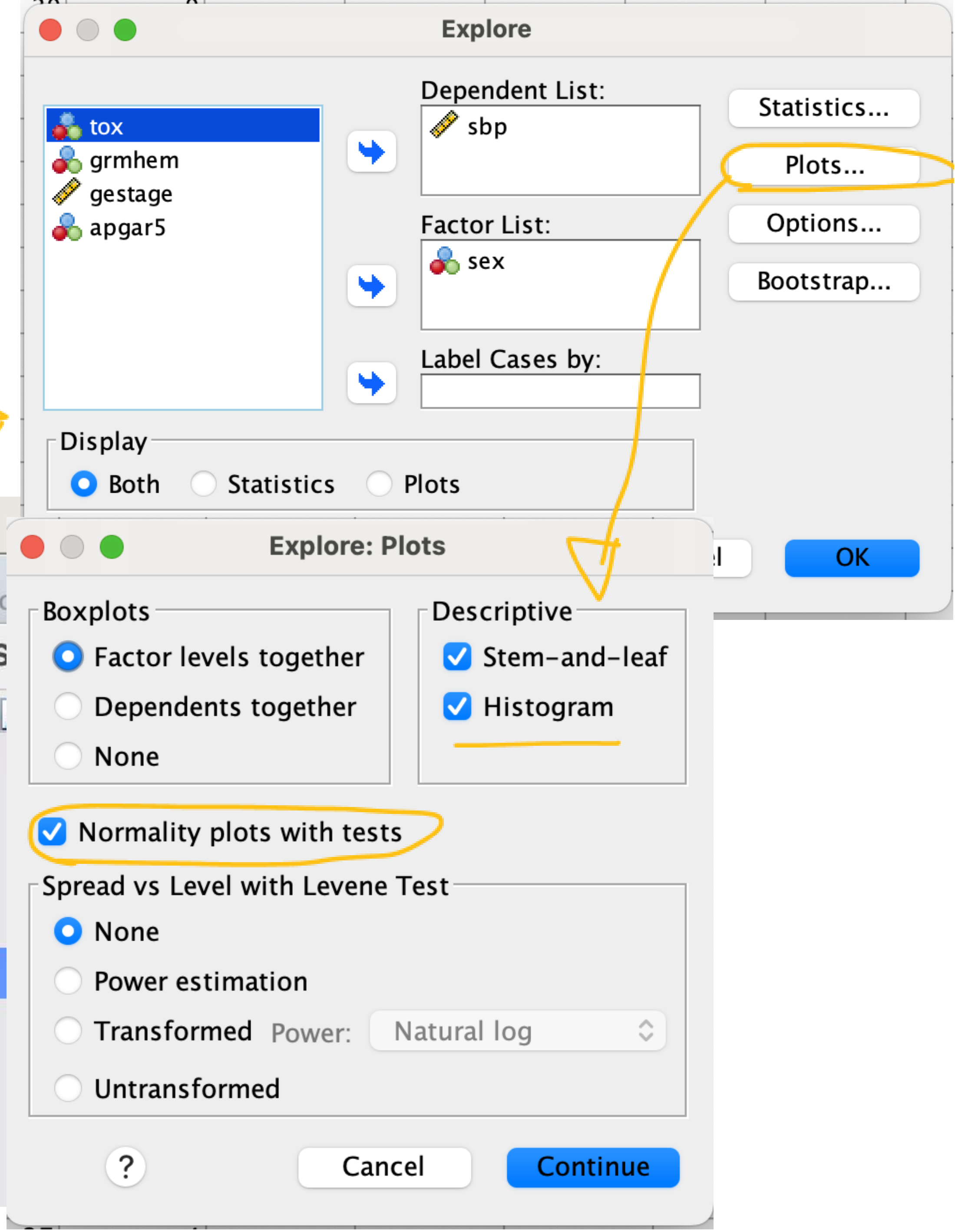
- The means of the two populations being compared should follow normal distributions. Under weak assumptions, this follows in large samples from the central limit theorem, even when the distribution of observations in each group is non-normal.[19] $n \geq 30$
- If using Student's original definition of the *t*-test, the two populations being compared should have the same variance (testable using *F*-test, Levene's test, Bartlett's test, or the Brown–Forsythe test; or assessable graphically using a Q–Q plot). If the sample sizes in the two groups being compared are equal, Student's original *t*-test is highly robust to the presence of unequal variances.[20] Welch's *t*-test is insensitive to equality of the variances regardless of whether the sample sizes are similar.
- The data used to carry out the test should either be sampled independently from the two populations being compared or be fully paired. This is in general not testable from the data, but if the data are known to be dependent (e.g. paired by test design), a dependent test has to be applied. For partially paired data, the classical independent *t*-tests may give invalid results as the test statistic might not follow a *t* distribution, while the dependent *t*-test is sub-optimal as it discards the unpaired data.[21]

Most two-sample *t*-tests are robust to all but large deviations from the assumptions.[22]

Describe data

Let’s also describe the data using numbers.





Normality check: test results

Tests of Normality

	sex	Kolmogorov-Smirnov[a] Statistic	df	Sig.	Shapiro-Wilk Statistic	df	Sig.
sbp	0	.091	56	.200*	.968	56	.146
	1	.143	44	.025	.938	44	.020

*. This is a lower bound of the true significance.

a. Lilliefors Significance Correction

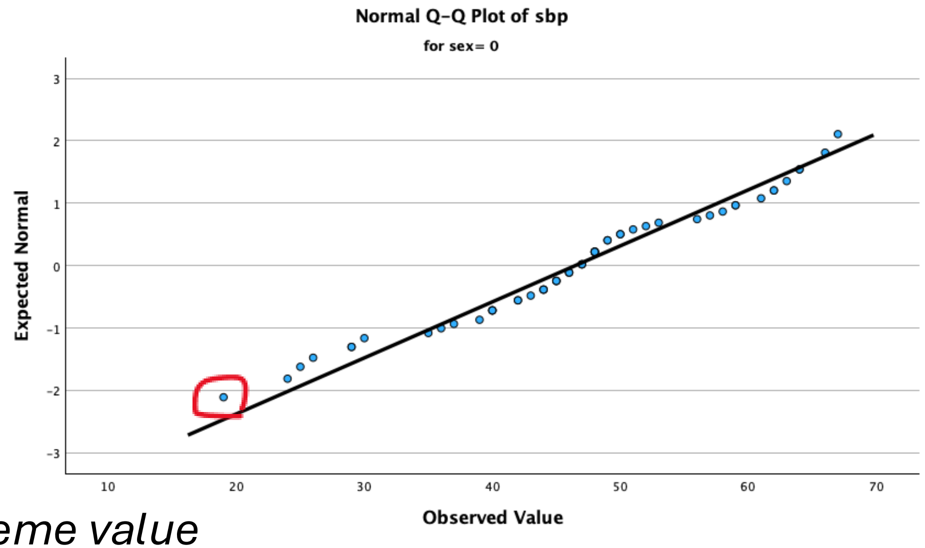
If the p-value of either test is < 0.05 (typical alpha level), the interpretation is: **normality assumption violated**

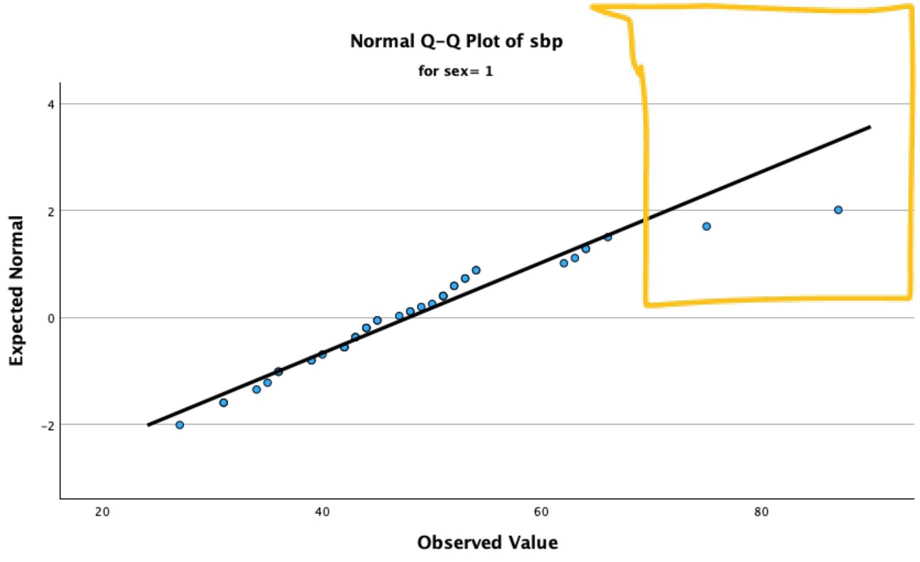
Do check other means to check the normality – tests are sensitive to outliers.

Normality check: Q-Q plot

- Can you use q-q plots to tell if data of each sample (approximately) follow a normal distribution?

Extreme values

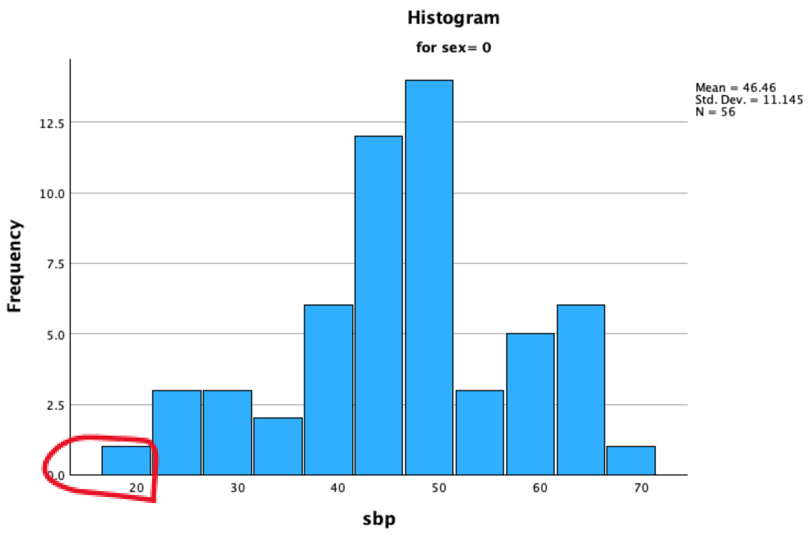


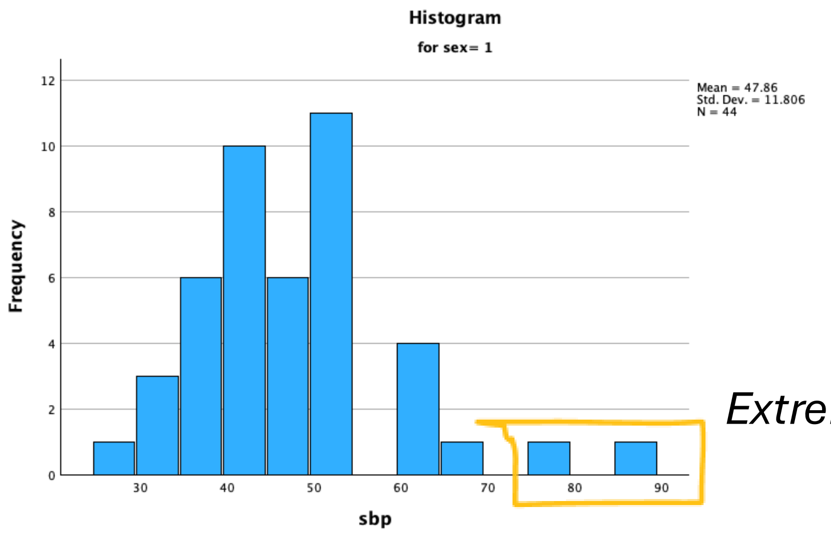


Extreme value

Describe data

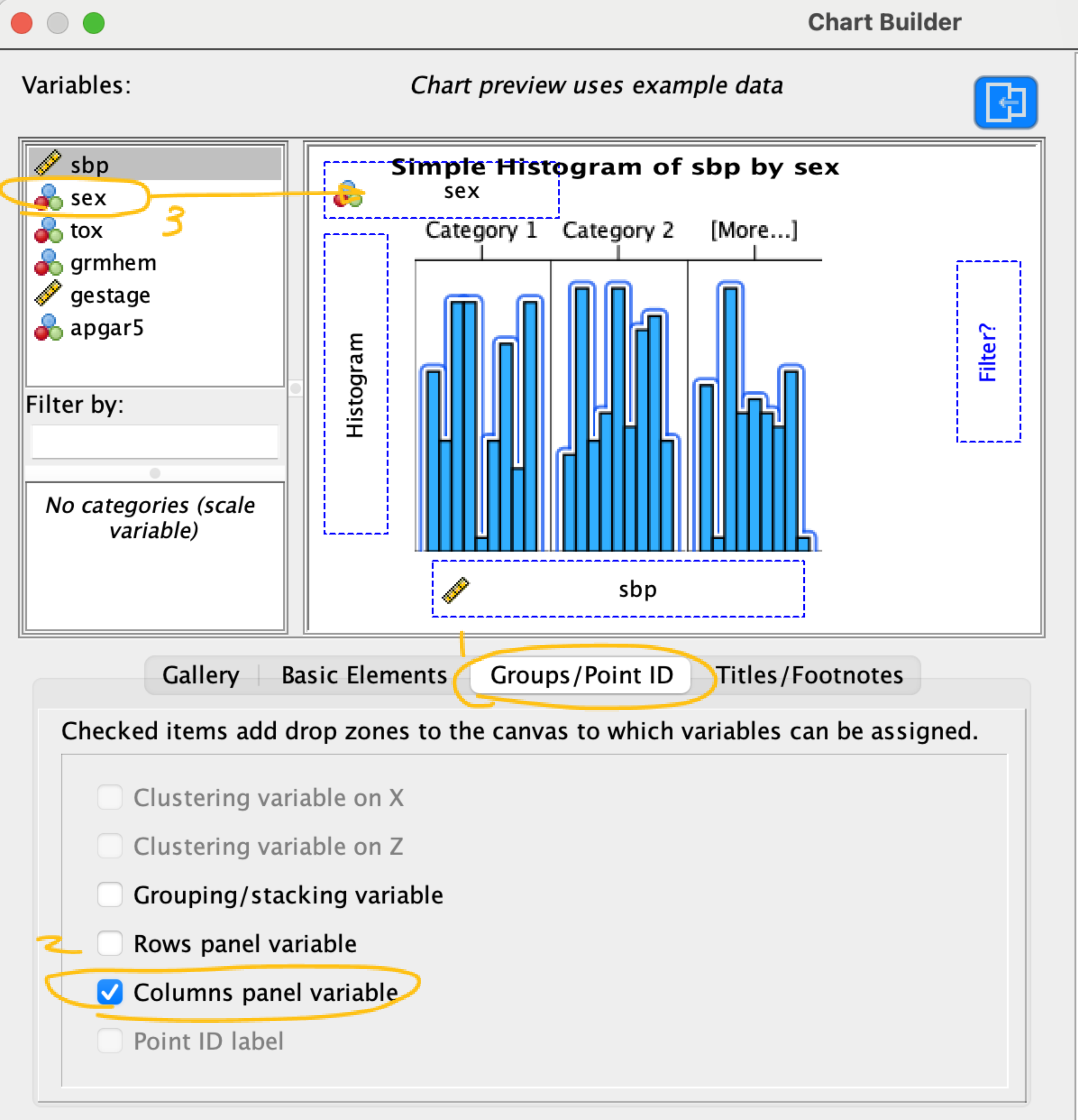
- Your potential audience may be more familiar with histograms



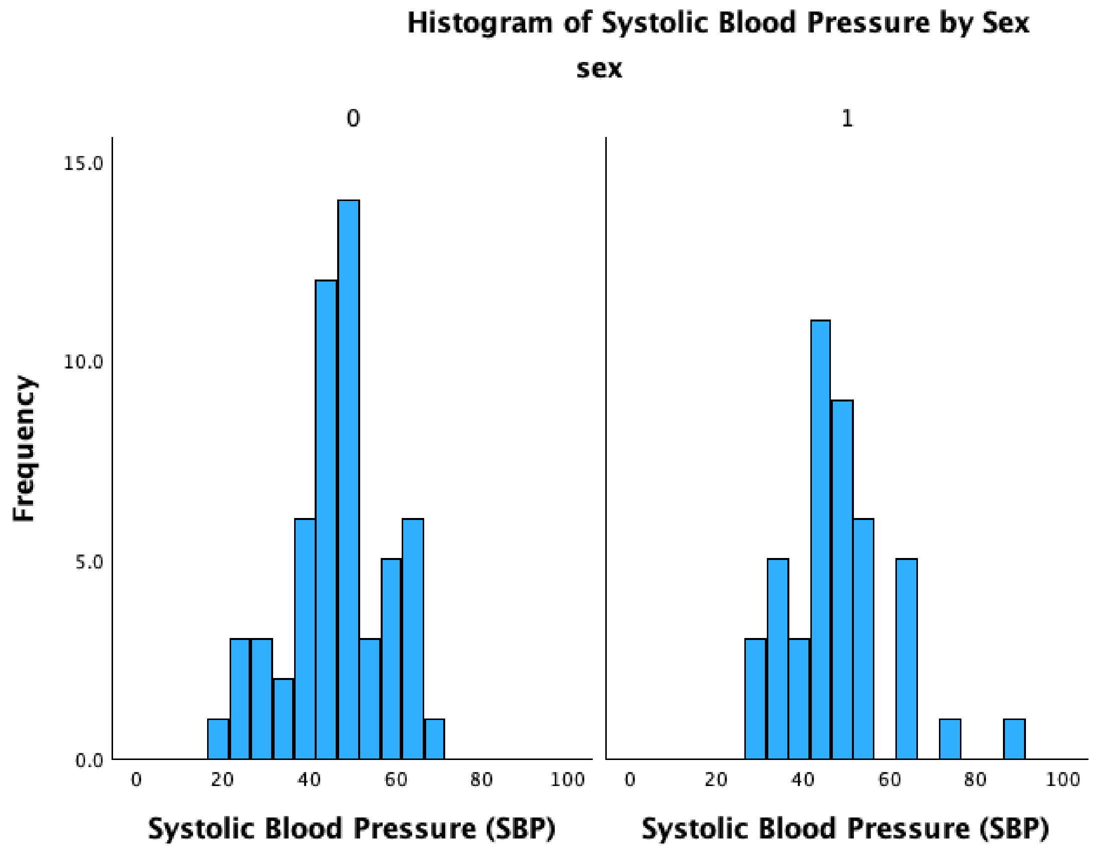


Extreme values

Extreme value



If you want both histograms in one figure



Describe data

- What's the sample size for sex=0 and sex=1?

- What's the variance/SD value of each sample? Are the two samples similar in the variance value?

Descriptives

	sex			Statistic	Std. Error
sbp	0	Mean		46.46	1.489
		95% Confidence Interval for Mean	Lower Bound	43.48	
			Upper Bound	49.45	
				46.72	
				47.00	
				124.217	
				11.145	
				19	
				67	
				48	
		Interquartile Range		13	
		Skewness		-.282	.319
		Kurtosis		-.071	.628
	1	Mean		47.86	1.780
		95% Confidence Interval for Mean	Lower Bound	44.27	
			Upper Bound	51.45	
		5% Trimmed Mean		47.14	
		Median		46.00	
		Variance		139.376	
		Std. Deviation		11.806	
		Minimum		27	
		Maximum		87	
		Range		60	
		Interquartile Range		12	
		Skewness		1.038	.357
		Kurtosis		1.909	.702

sex

Case Processing Summary

		Cases					
		Valid		Missing		Total	
	sex	N	Percent	N	Percent	N	Percent
sbp	0	56	100.0%	0	0.0%	56	100.0%
	1	44	100.0%	0	0.0%	44	100.0%

You can run independent sample t-test

- Sample distribution of Sbp when sex = 1 is not following a normal distribution, mostly because of the two outliers. The sample size of n=44 also supports the idea that the *sampling distribution* of the sample mean will be a normal distribution (central limit theorem).

- Proceed with conducting the t-test

T-test results

In the *t*-test comparing the means of two independent samples, the following assumptions should be met:

- The means of the two populations being compared should follow normal distributions. Under weak assumptions, this follows in large samples from the central limit theorem, even when the distribution of observations in each group is non-normal.[19]
- If using Student's original definition of the *t*-test, the two populations being compared should have the same variance (testable using *F*-test, Levene's test, Bartlett's test, or the Brown–Forsythe test; or assessable graphically using a Q–Q plot). If the sample sizes in the two groups being compared are equal, Student's original *t*-test is highly robust to the presence of unequal variances.[20] Welch's *t*-test is insensitive to equality of the variances regardless of whether the sample sizes are similar.
- The data used to carry out the test should either be sampled independently from the two populations being compared or be fully paired. This is in general not testable from the data, but if the data are known to be dependent (e.g. paired by test design), a dependent test has to be applied. For partially paired data, the classical independent *t*-tests may give invalid results as the test statistic might not follow a *t* distribution, while the dependent *t*-test is sub-optimal as it discards the unpaired data.[21]

Most two-sample *t*-tests are robust to all but large deviations from the assumptions.[22]

T-Test

Group Statistics

	sex	N	Mean	Std. Deviation	Std. Error Mean
sbp	0	56	46.46	11.145	1.489
	1	44	47.86	11.806	1.780

Independent Samples Test

		Levene's Test for Equality of Variances		t-test for Equality of Means					
						Significance			
		F	Sig.	t	df	One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference
sbp	Equal variances assumed	.079	.779	-.607	98	.273	.545	-1.399	2.305
	Equal variances not assumed			-.603	89.858	.274	.548	-1.399	2.321

p < 0.05 (typical alpha level) indicates equal variance assumption violated

T-test results: effect size

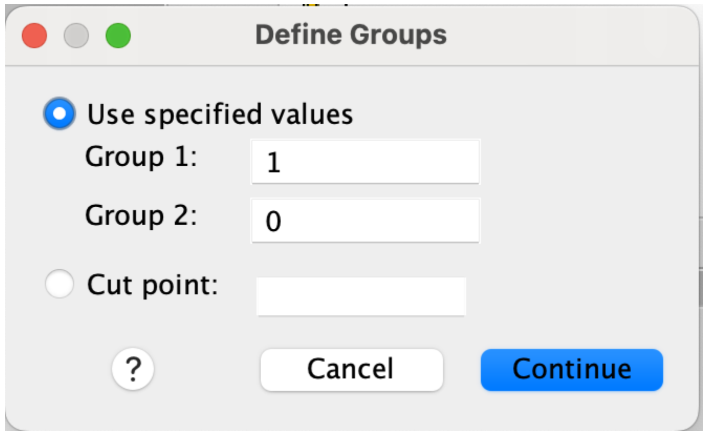
Independent Samples Effect Sizes

		Standardizer[a]	Point Estimate	95% Confidence Interval Lower	95% Confidence Interval Upper
sbp	Cohen's d	11.440	-.122	-.517	.273
	Hedges' correction	11.528	-.121	-.513	.271
	Glass's delta	11.806	-.119	-.513	.278

a. The denominator used in estimating the effect sizes.
Cohen's d uses the pooled standard deviation.
Hedges' correction uses the pooled standard deviation, plus a correction factor.
Glass's delta uses the sample standard deviation of the control (i.e., the second) group.

Most common statistic reported. You can report the absolute value of the point estimate (again, that's more common). In doing so, confidence intervals also need to be 'flipped' (ex. -.273 ~ .517.

You can try checking the above statement by conducting t-test again with different group levels.

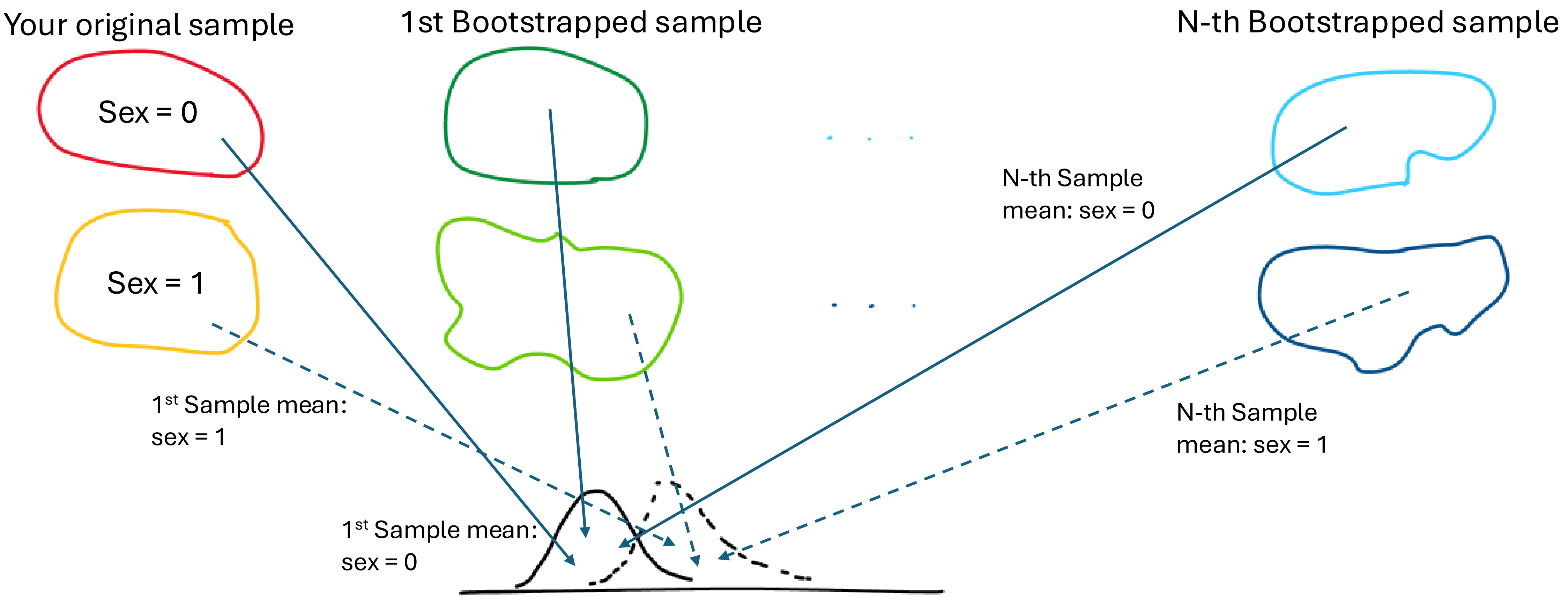


Practice reporting results

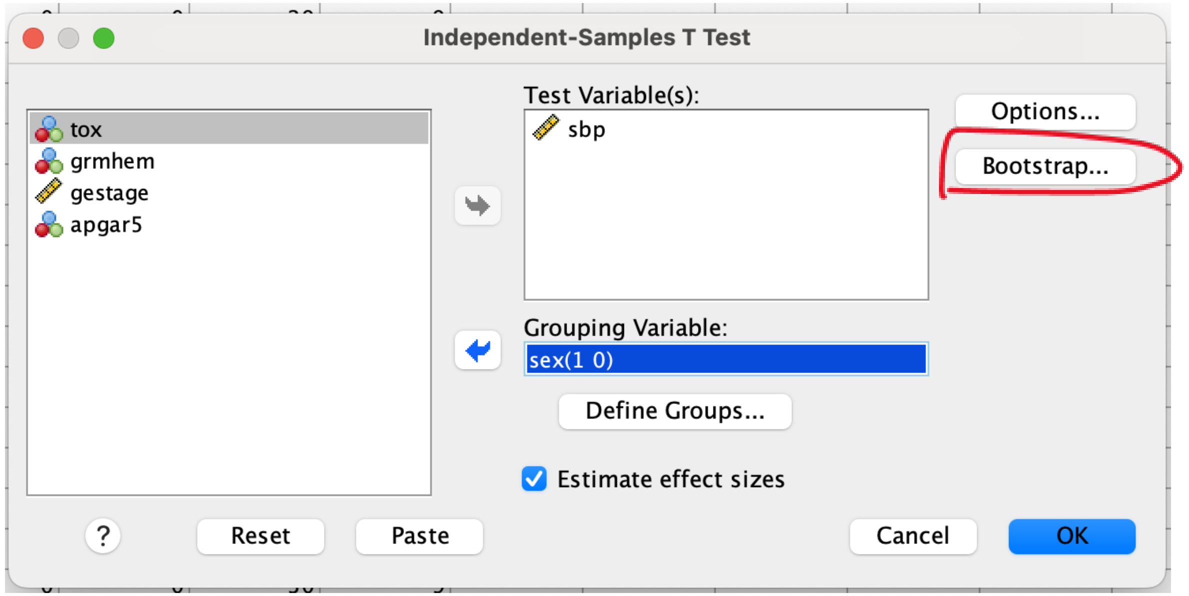
- Please write a paragraph to report the statistical analysis results you generated so far.

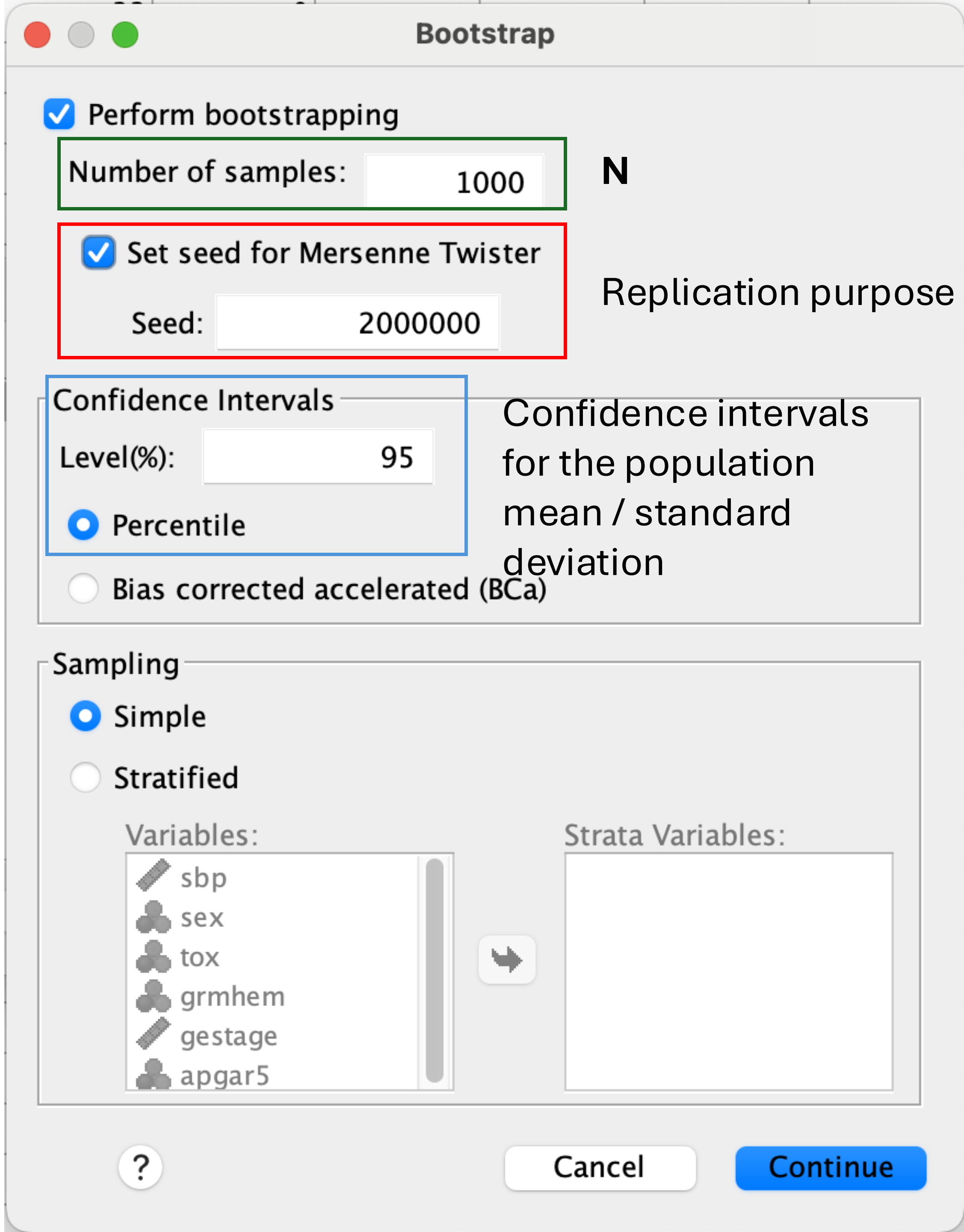
Extra: bootstrapping

- You can *simulate* the **sampling distributions** by bootstrapping.



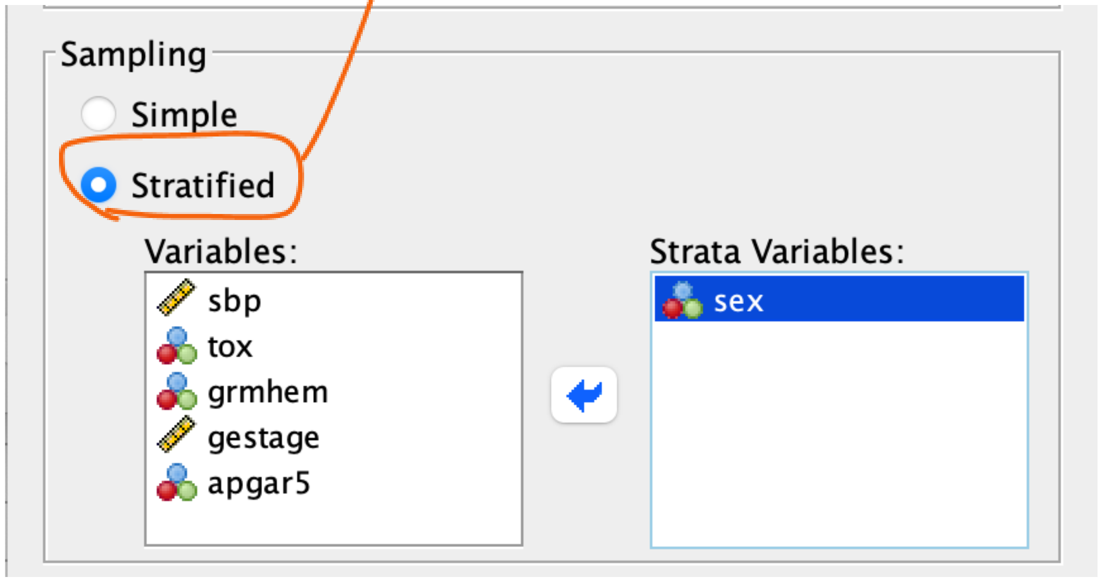
Extra: bootstrapping

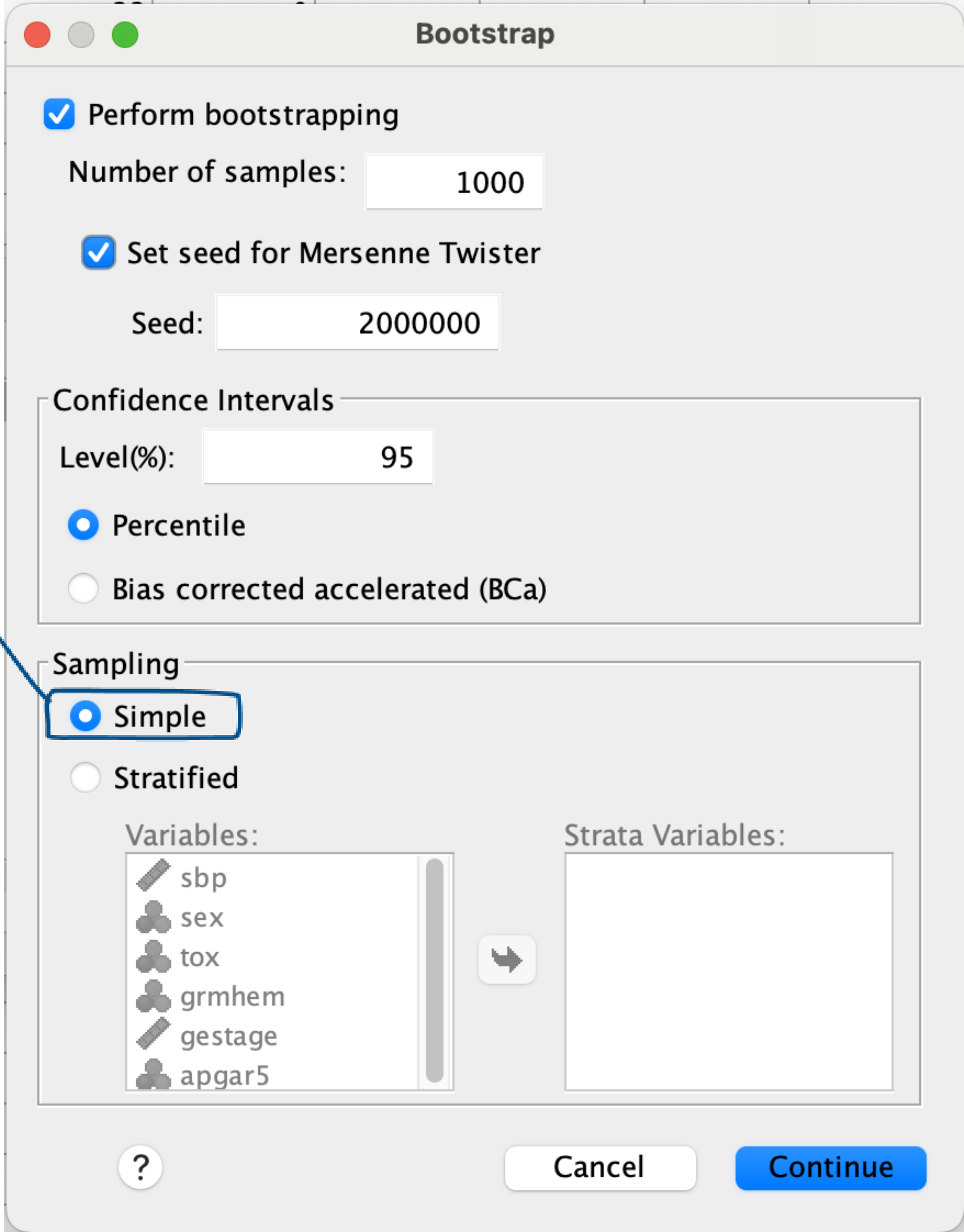




Extra: bootstrapping

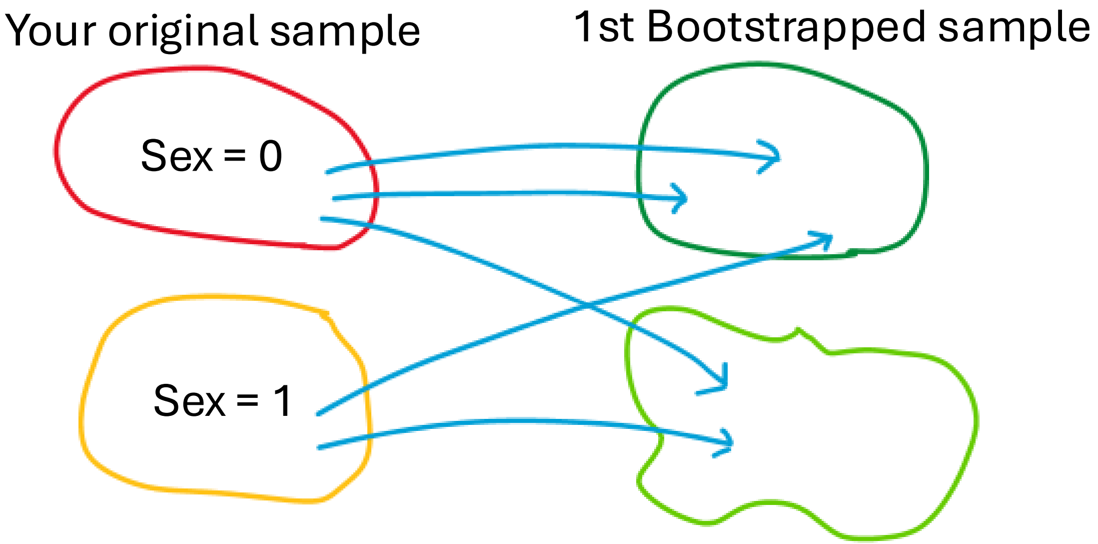
If "Simple", resample from the entire sample
If "Stratified", resample with respect to strata variable(s)



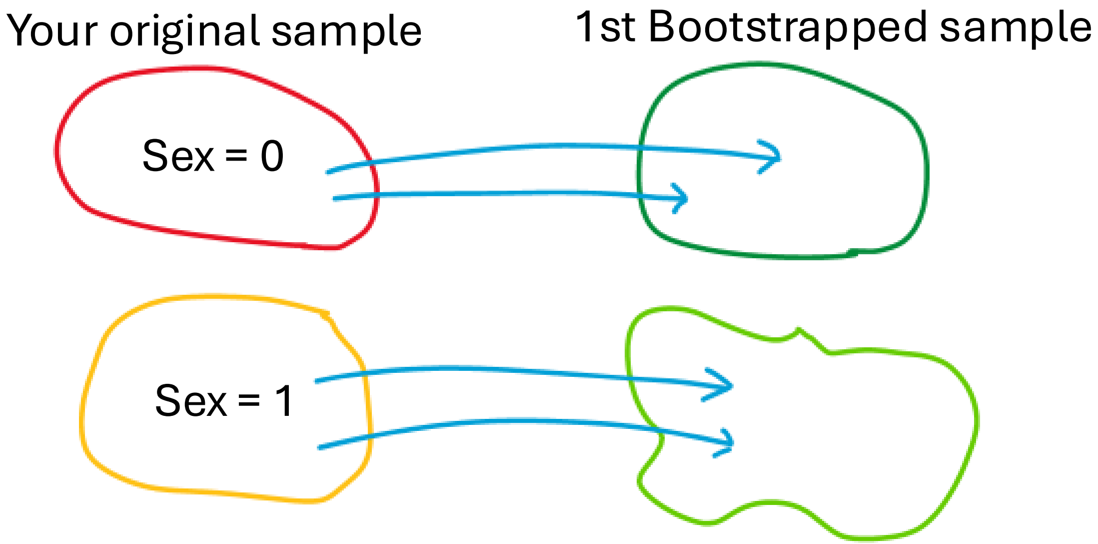


Extra: bootstrapping

- Simple vs. Stratified re-sampling



Simple: does not account for the strata variable (*sex*); the new sample's *sbp* values for *sex* = 0 can be resampled from the original sample's *sbp* values of either *sex* = 0 or *sex* = 1.



Stratified: maintains the original distribution of the strata variable (*sex*); the new sample's *sbp* values for *sex* = 1 are resampled only from the original sample's *sbp* values of *sex* = 1, and the same applies for *sex* = 0.

Extra: bootstrapping

Bootstrap

Sampling = Simple

Bootstrap Specifications

Sampling Method	Simple
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile

T-Test

Group Statistics

	sex		Statistic	Bootstrap[a] Bias	Std. Error	95% Confidence Interval Lower	Upper
sbp	1	N	44				
		Mean	47.86	-.02	1.78	44.54	51.52
		Std. Deviation	11.806	-.211	1.650	8.427	14.950
		Std. Error Mean	1.780				
	0	N	56				
		Mean	46.46	.00	1.53	43.35	49.33
		Std. Deviation	11.145	-.151	.995	8.966	12.801
		Std. Error Mean	1.489				

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Bootstrap

Sampling = Stratified by Sex

Bootstrap Specifications

Sampling Method	Stratified
Number of Samples	1000
Confidence Interval Level	95.0%
Confidence Interval Type	Percentile
Strata Variables	sex

T-Test

Group Statistics

	sex		Statistic	Bootstrap[a] Bias	Std. Error	95% Confidence Interval Lower	Upper
sbp	1	N	44				
		Mean	47.86	-.04	1.78	44.55	51.52
		Std. Deviation	11.806	-.224	1.673	8.498	14.827
		Std. Error Mean	1.780				
	0	N	56				
		Mean	46.46	.03	1.50	43.66	49.50
		Std. Deviation	11.145	-.160	1.008	8.969	12.869
		Std. Error Mean	1.489				

a. Unless otherwise noted, bootstrap results are based on 1000 stratified bootstrap samples

Sampling = “Stratified”

Bootstrap for Independent Samples Test

		Mean Difference	Bootstrap[a]				
			Bias	Std. Error	Sig. (2-tailed)	95% Confidence Interval Lower	Upper
sbp	Equal variances assumed	1.399	-.071	2.311	.557	-3.111	5.851
	Equal variances not assumed	1.399	-.071	2.311	.557	-3.111	5.851

a. Unless otherwise noted, bootstrap results are based on 1000 stratified bootstrap samples

Bootstrap for Independent Samples Test

Sampling = “Simple”

		Mean Difference	Bootstrap[a]			
			Bias	Std. Error	95% Confidence Interval Lower	Upper
sbp	Equal variances assumed	1.399	-.024	2.325	-3.139	5.908
	Equal variances not assumed	1.399	-.024	2.325	-3.139	5.908

a. Unless otherwise noted, bootstrap results are based on 1000 bootstrap samples

Independent Samples Test

Parametric test results

	t-test for Equality of Means							
	t	df	Significance One-Sided p	Two-Sided p	Mean Difference	Std. Error Difference	95% Confidence Interval of the Difference Lower	Upper
)	.607	98	.273	.545	1.399	2.305	-3.174	5.973
	.603	89.858	.274	.548	1.399	2.321	-3.211	6.010

Practice reporting results

- Please write a paragraph to report the statistical analysis results you generated using bootstrapping method (either simple or stratified method).