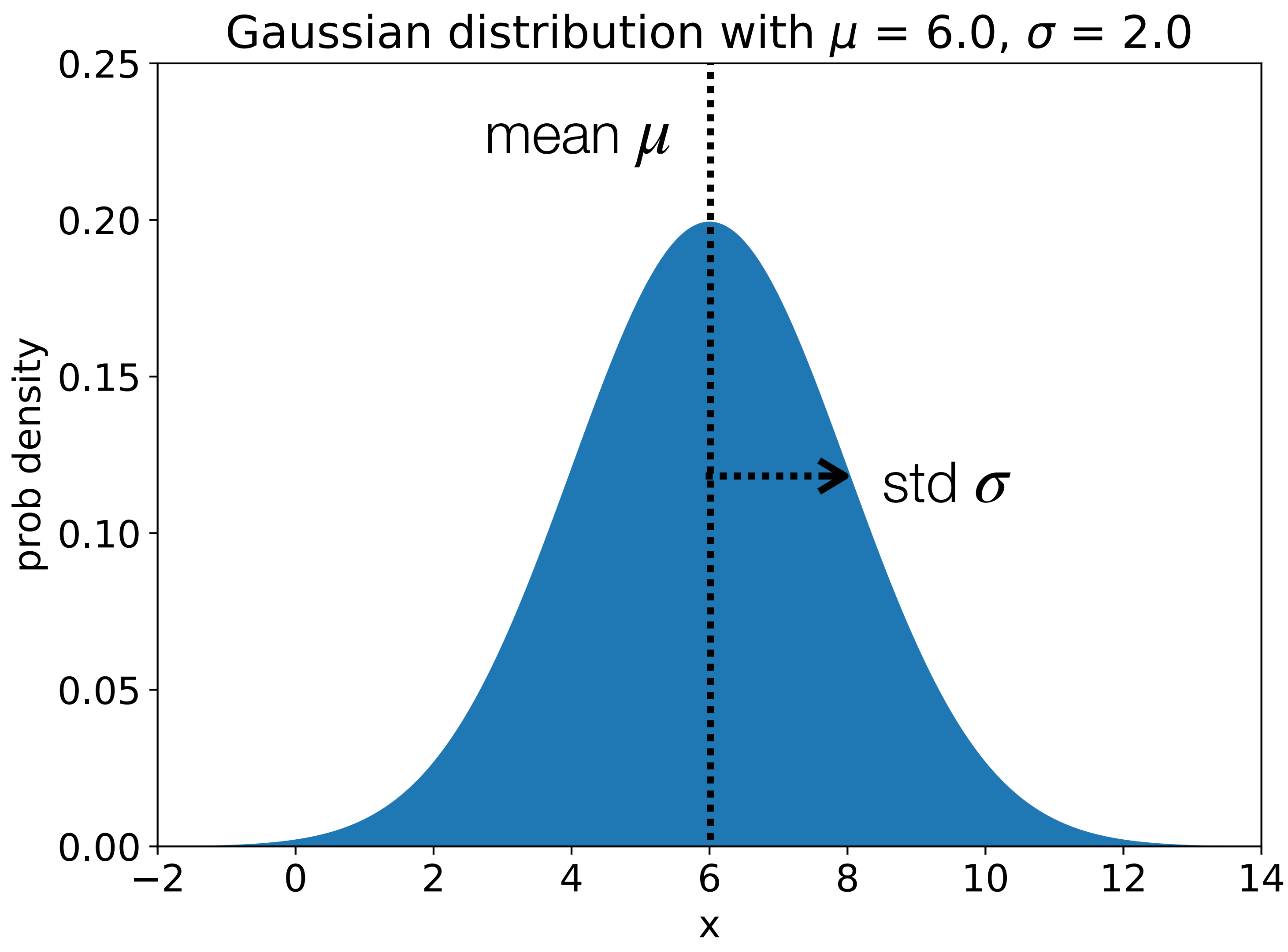
Gaussian distributions
QQ plots
t-tests
Comparing two datasets

Biostatistics Course 2023
Lecture 3
Wednesday, 26 July 2020
1:00pm - 3:00pm

Gaussian distributions

The normal distribution is ubiquitous in statistics

"Gaussian distribution" = "normal distribution"



$$x \sim \mathrm{Normal}(\mu, \sigma^2)$$

drawn from

mean

variance

Mean and variance

Let X ~ N(μ, σ²)

- Mean: $E[X] = \mu$
- Variance: $Var[X] = \sigma^2$
- Standard Deviation: $SD_X = \sigma$

Mean of standardized random variable

Let

$$Z = (Y - \mu)/\sigma$$

$$\begin{aligned} E[Z] = E\left[\frac{Y-\mu}{\sigma}\right] &= \frac{1}{\sigma}E[Y-\mu] \\ &= \frac{1}{\sigma}(E[Y]-\mu) \\ &= \frac{1}{\sigma}(\mu-\mu) \\ &= 0 \end{aligned}$$

Variance of standardized random variable

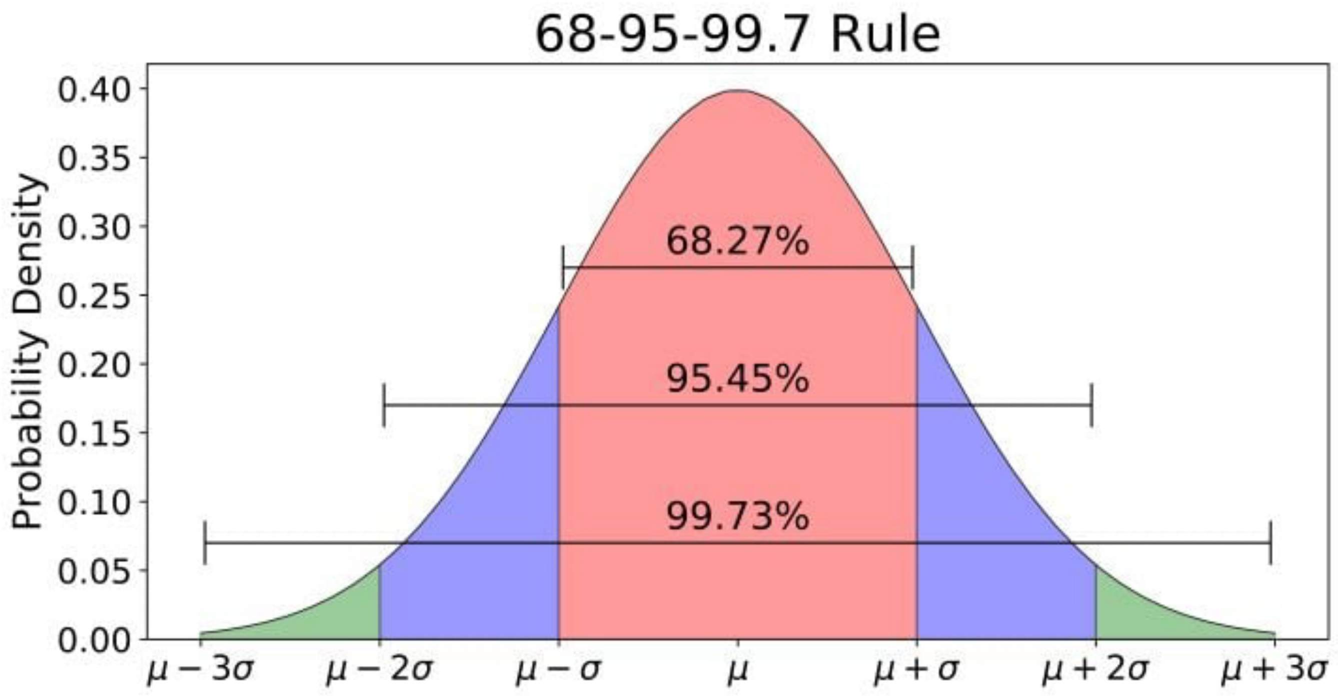
$$\begin{aligned} Var[Z] &= Var\left[\frac{Y-\mu}{\sigma}\right] \\ &= \frac{1}{\sigma^2} Var[Y-\mu] \\ &= \frac{1}{\sigma^2} Var[Y] \\ &= \frac{1}{\sigma}\sigma^2 \\ &= 1 \end{aligned}$$

NOTE: $\mu = 0$ and $\sigma^2 = 1$ for **any** standardized random variable

68-95-99.7 Rule

Recall the 68-95-99.7 rule Note for a standard normal random variable, Z ~ N(0, 1)

$$Pr(-1 < Z < 1) \approx 0.68$$
$$Pr(-2 < Z < 2) \approx 0.95$$
$$Pr(-3 < Z < 3) \approx 0.997$$



The central limit theorem makes the normal distribution extremely relevant

If a random variable X has population mean μ and population variance σ^2, the sample mean X-bar, based on n observations, is approximately normally distributed with mean μ and variance σ^2, for sufficiently large n.

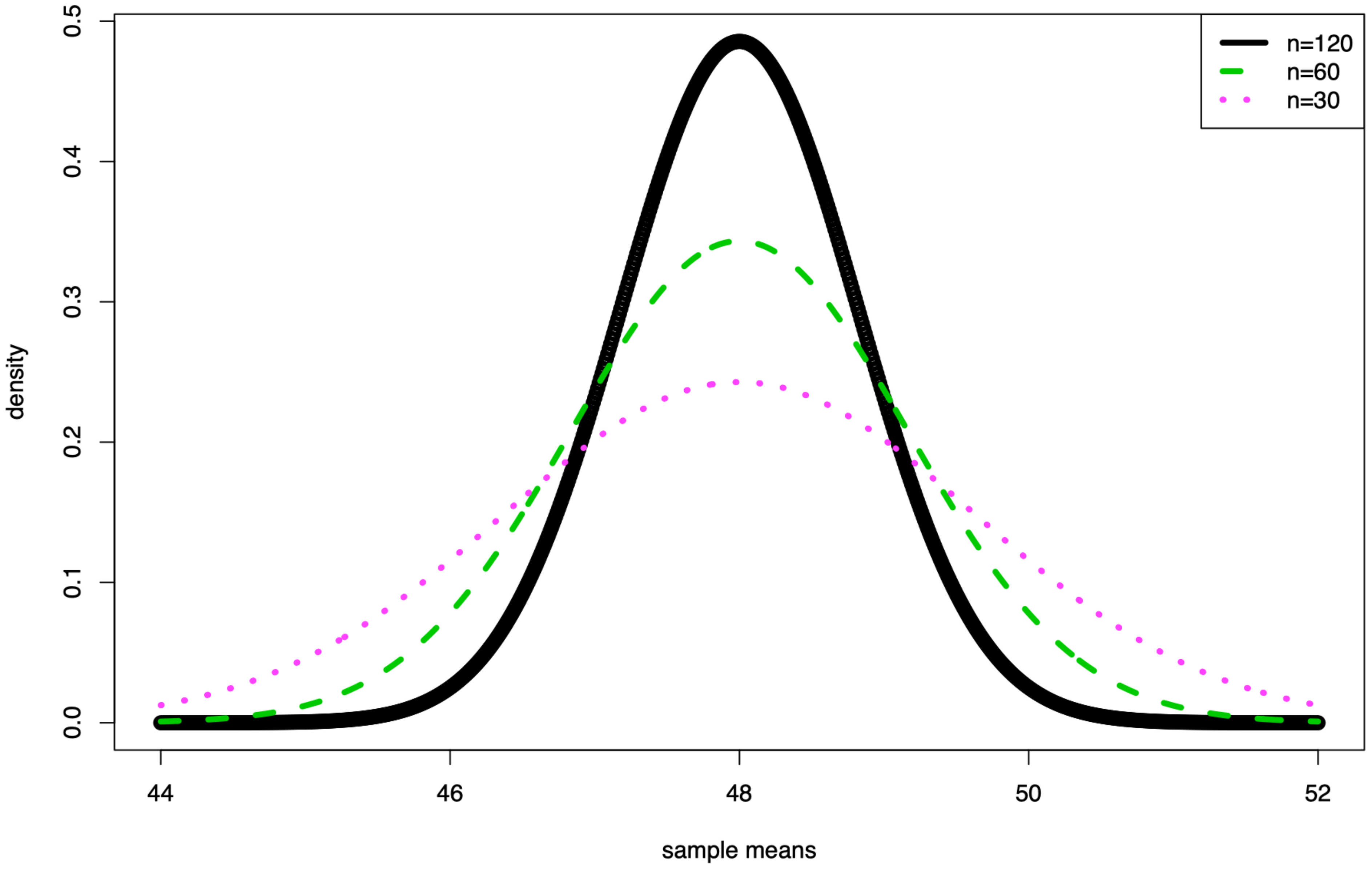
$$
\begin{aligned}
x_1 &\sim p_1(x) \\
x_2 &\sim p_2(x) \\
&\cdots \\
x_N &\sim p_N(x)
\end{aligned}
\qquad
\bar{x} = \frac{x_1 + x_2 + \cdots + x_N}{N}
\quad \Rightarrow \quad
\bar{x} \sim \text{Normal}(\mu, \sigma^2)
$$

This means that, if many sources additively contribute to an experimental measurement, independent measurements will be approximately normally distributed.

This is why statisticians so often assume that experimental measurements follow normal distributions.

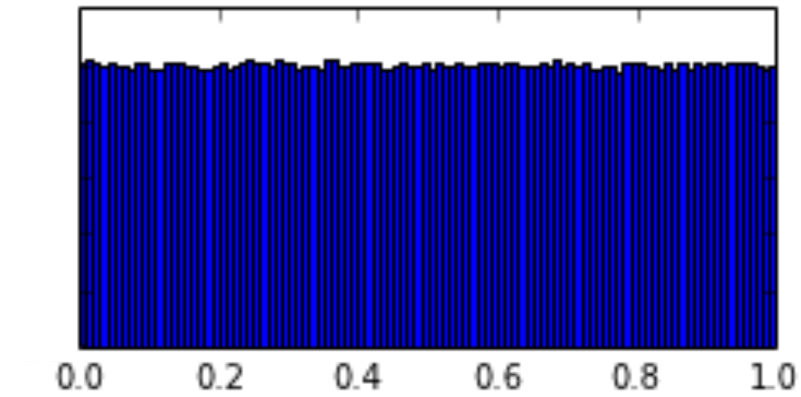
Impact of sample size on sampling distribution

Sample 1 (n=30); sample 2 (n=60); sample 3 (n=120)

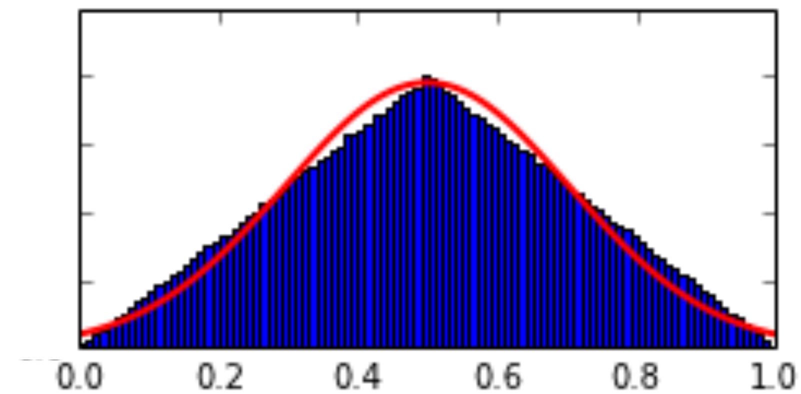


Suppose $x_1, x_2, \cdots, x_N$ are drawn from a uniform (i.e. flat) probability distribution that stands from 0 and 1

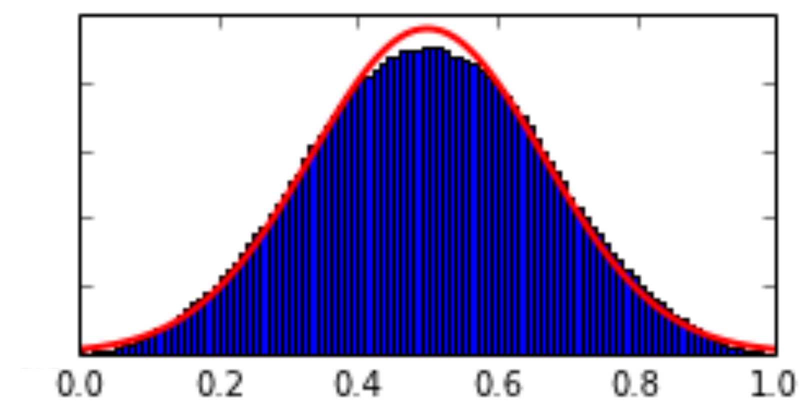
$$x_1$$



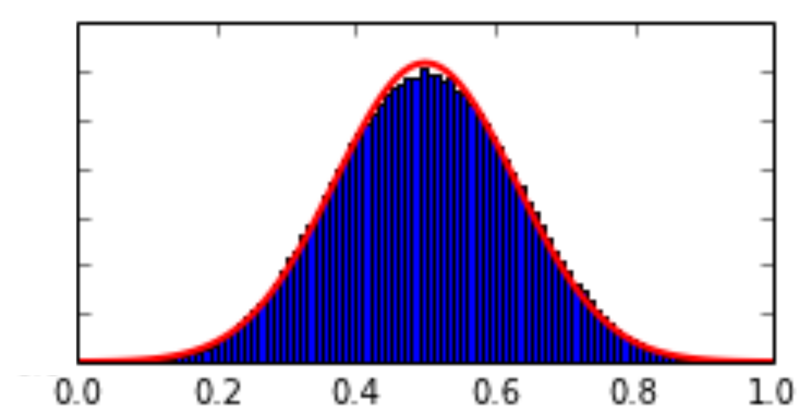
$$\frac{x_1 + x_2}{2}$$



$$\frac{x_1 + x_2 + x_3}{3}$$



$$\frac{x_1 + x_2 + x_3 + x_4 + x_5}{5}$$



Example 1: Human Sex Ratio

The human sex ratio at birth is slightly skewed towards boys rather than girls.

	count
male	484382
female	453841
total	938223

→

probability of male birth

estimate: 51.63%

95% CI: [51.53%, 51.73%]

Arbuthnot J (1711). An Argument for Divine Providence, taken from the Constant Regularity observed in the Births of both Sexes.

We assume the number of male babies (versus female babies) is drawn from a binomial distribution

data

n = 484,382: number of male births

N = 938,223: total number of births

model

$$n \sim \text{Binom}(q, N)$$

q: probity of a male birth

The assume probability distribution is called the <u>sampling distribution</u>

goals

1. Compute a <u>best estimate</u> $\hat{q}$ for q

2. Compute a <u>confidence interval</u> for q

The standard estimate of probability is just the ratio of counts

$n = 484{,}382$: number of male births

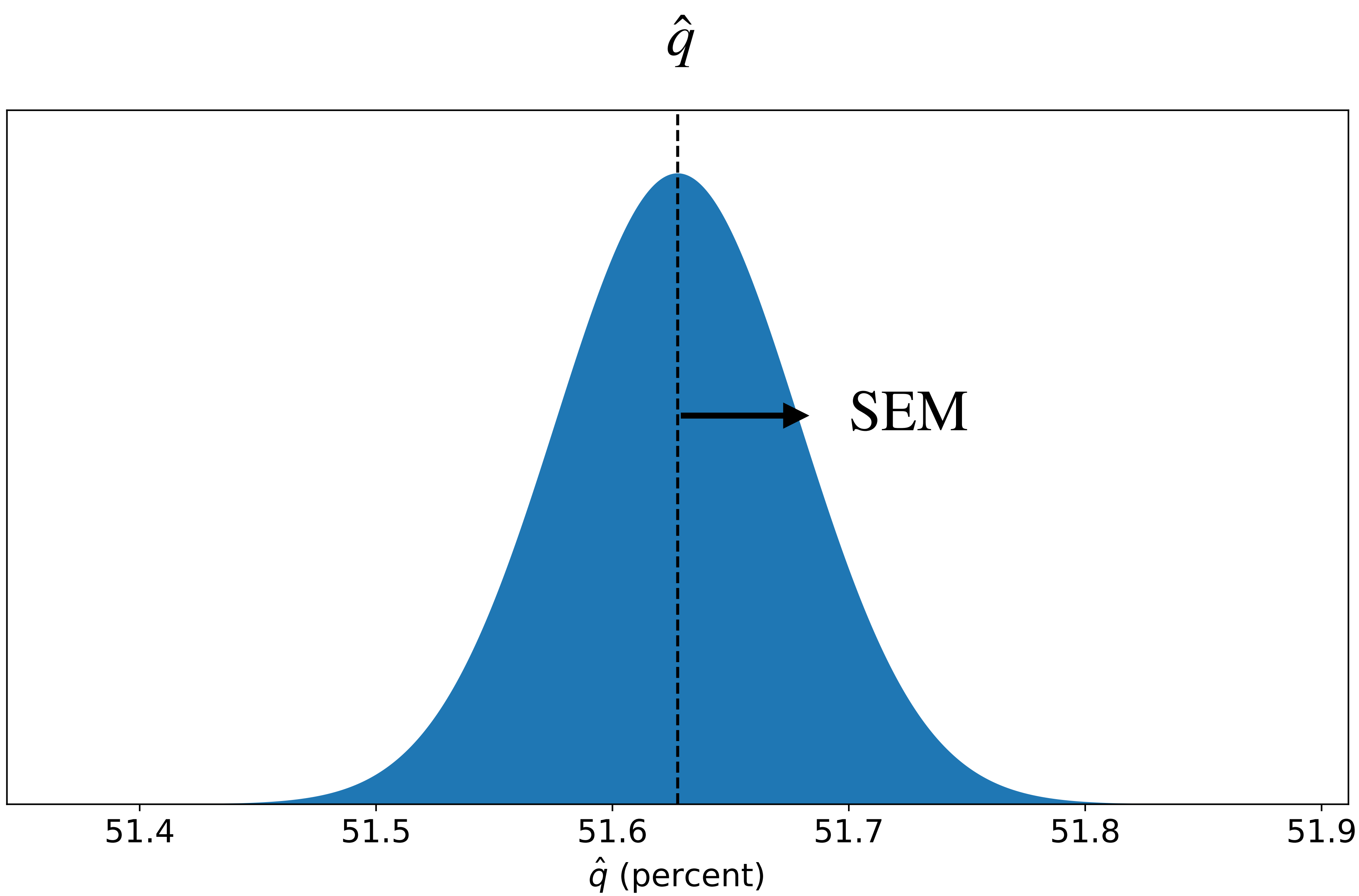
$N = 938{,}223$: total number of male births

$\hat{q} = \dfrac{n}{N} = 51.63\,\%$: estimated probability of a newborn being male

The lingering uncertainty in q is (verly nearly) described by a normal distribution centered on the estimate $\hat{q}$.

The standard deviation of this distribution is called the standard error of the mean (SEM).

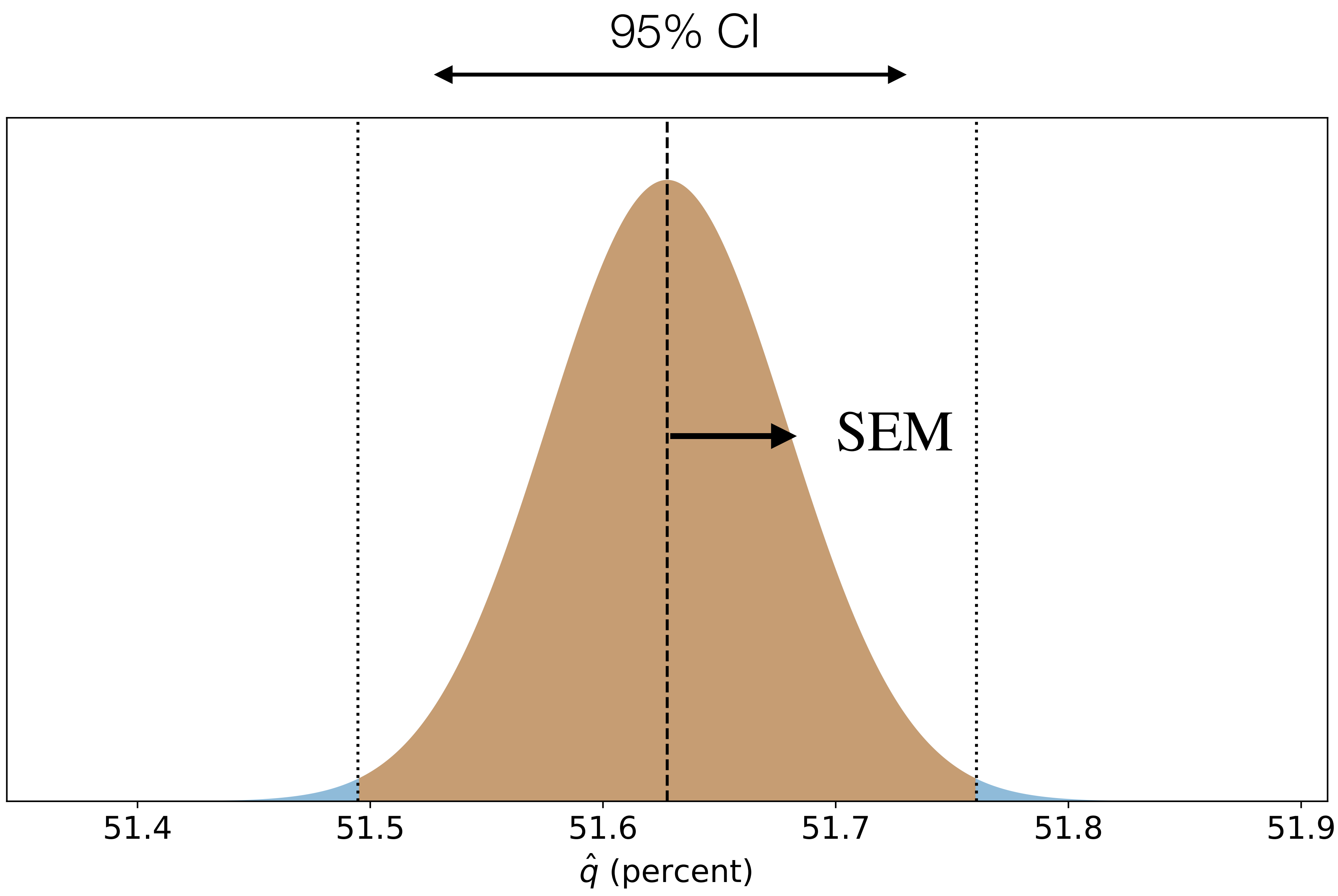
$$\mathrm{SEM} = \sqrt{\hat{q}(1-\hat{q})/N}$$



The 95% confidence interval, describing plausible values of q, is computed using both $\hat{q}$ and SEM.

The corresponding 95% confidence interval (CI) is

$$[\hat{q} - W, \hat{q} + W] \quad \text{where} \quad W = 1.96 \times \mathrm{SEM}$$



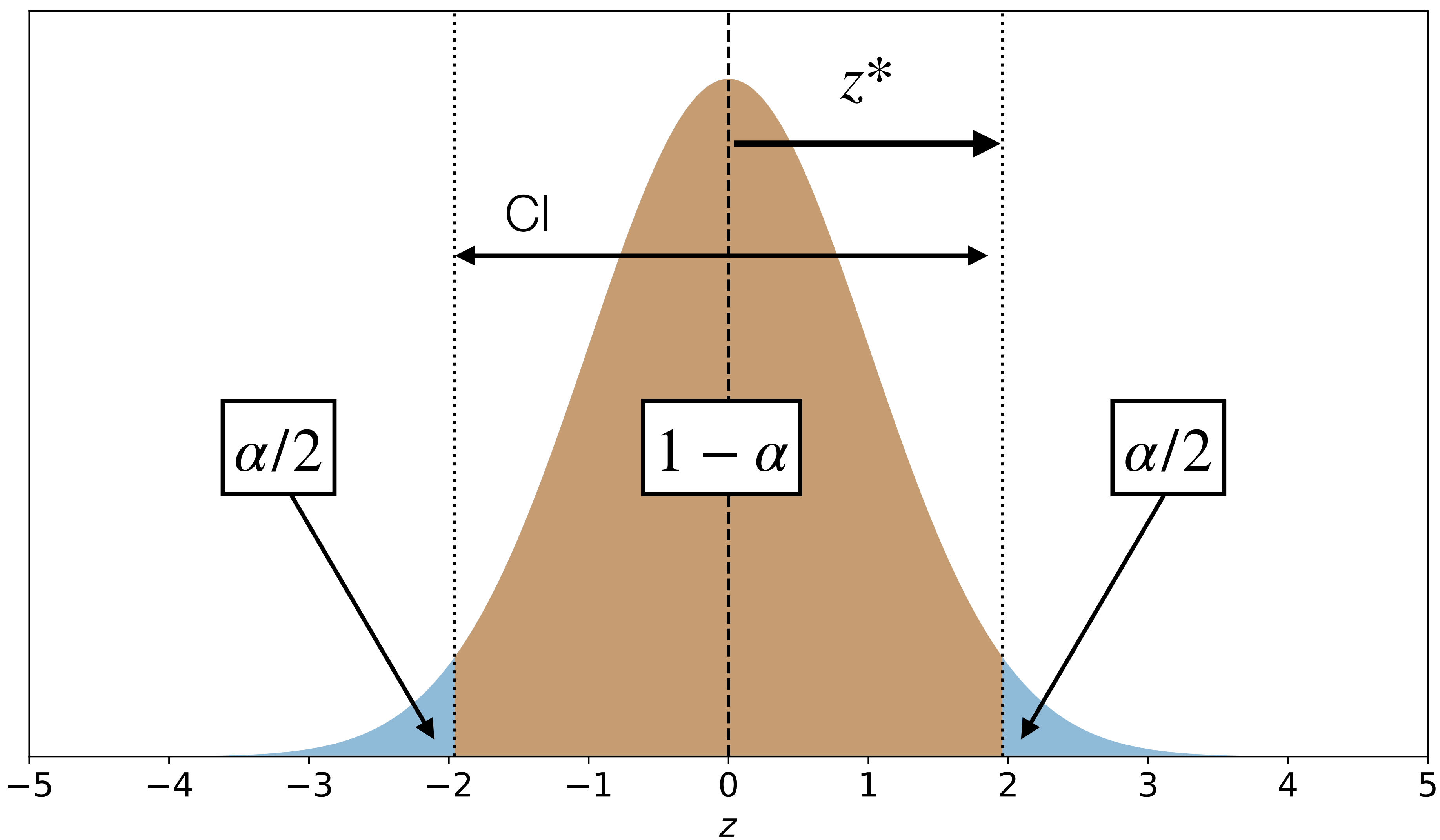
Uncertainty in q is summarized by a z-statistic

The z-statistic is defined by: $z = \dfrac{q - \hat{q}}{\mathrm{SEM}}$

Because of the central limit theorem, $z \sim \mathrm{Normal}(0,\ 1)$.

The user chooses a value for α, the probability that q is not within the confidence interval.

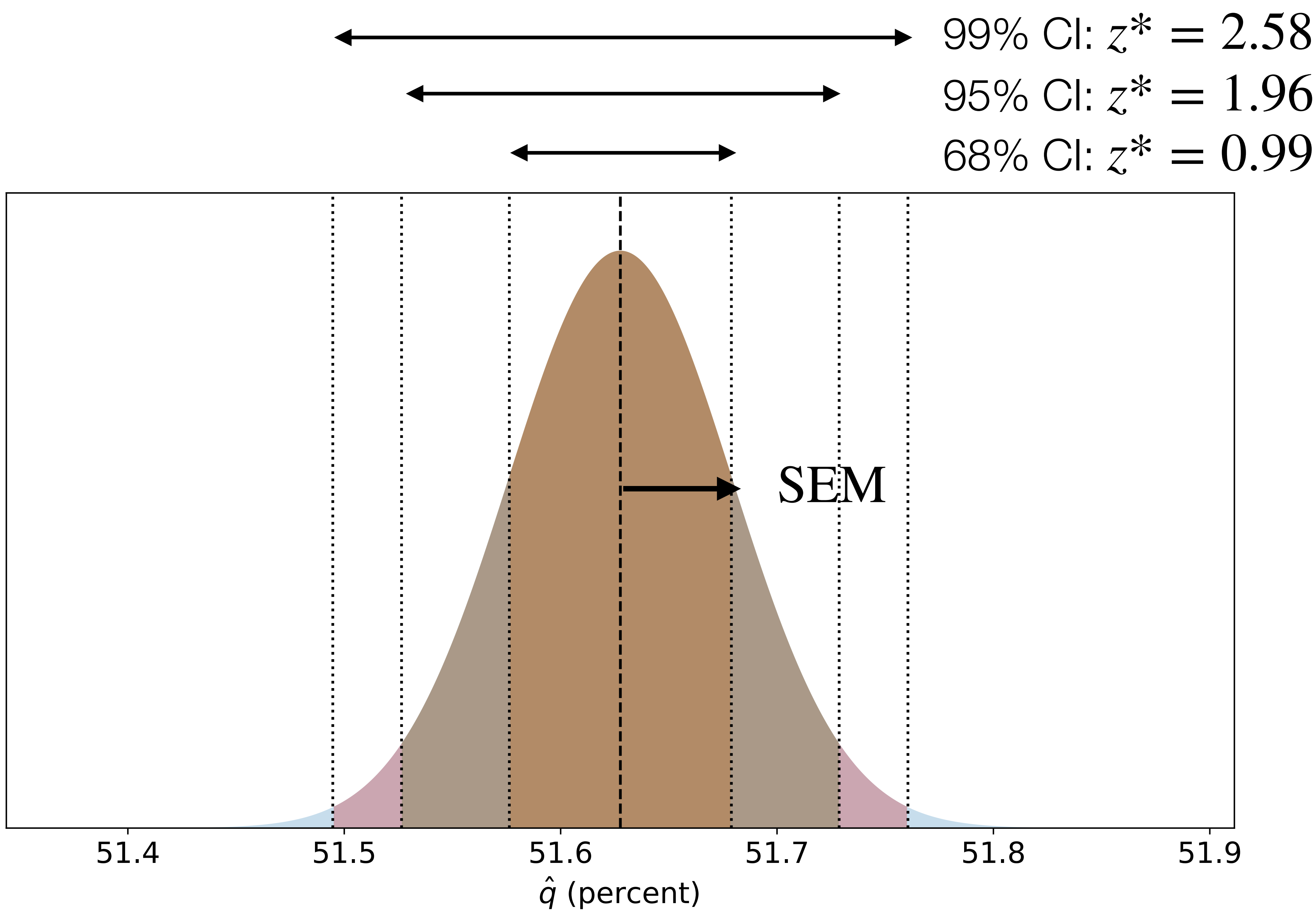
Choosing α fixes the value of z^*. Using $\alpha = 5\,\%$ gives $z^* = 1.96$.



Confidence intervals of different stringency can be computed using different z-statistic thresholds

Other confidence intervals are given by $[\hat{q} - W,\ \hat{q} + W]$ where

margin of error: $W = z^* \times \text{SEM}$



Example 2: Healthy Human Body Temperature

Example 2: Human body temperature

Body Temp	Sex	Heart Rate
96.3	2	70
96.7	2	71
96.9	2	74
97.0	2	80
97.1	2	73
97.1	2	75
97.1	2	82
97.2	2	64
97.3	2	69
97.4	2	70

⋮

Mackowiak PA, Wasserman SS, Levine MM. (1992) A Critical Appraisal of 98.6°F, the Upper Limit of the Normal Body Temperature, and Other Legacies of Carl Reinhold August Wunderlich. *JAMA.* 268(12):1578–1580.

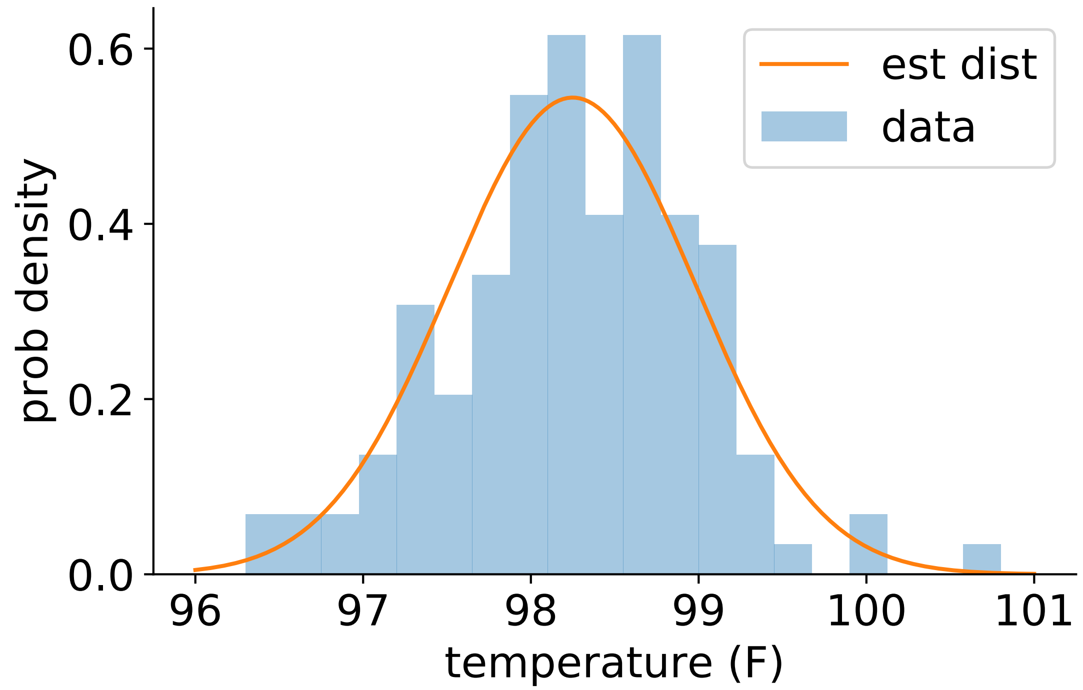
(Sex: 1 = female, 2 = male)

Example 2: Human Body Temperature

We model temperature using a normal distribution

Body Temp
96.3
96.7
96.9
97.0
97.1
97.1
97.1
97.2
97.3
97.4

⋮



temperature mean μ

estimate: 98.25 F

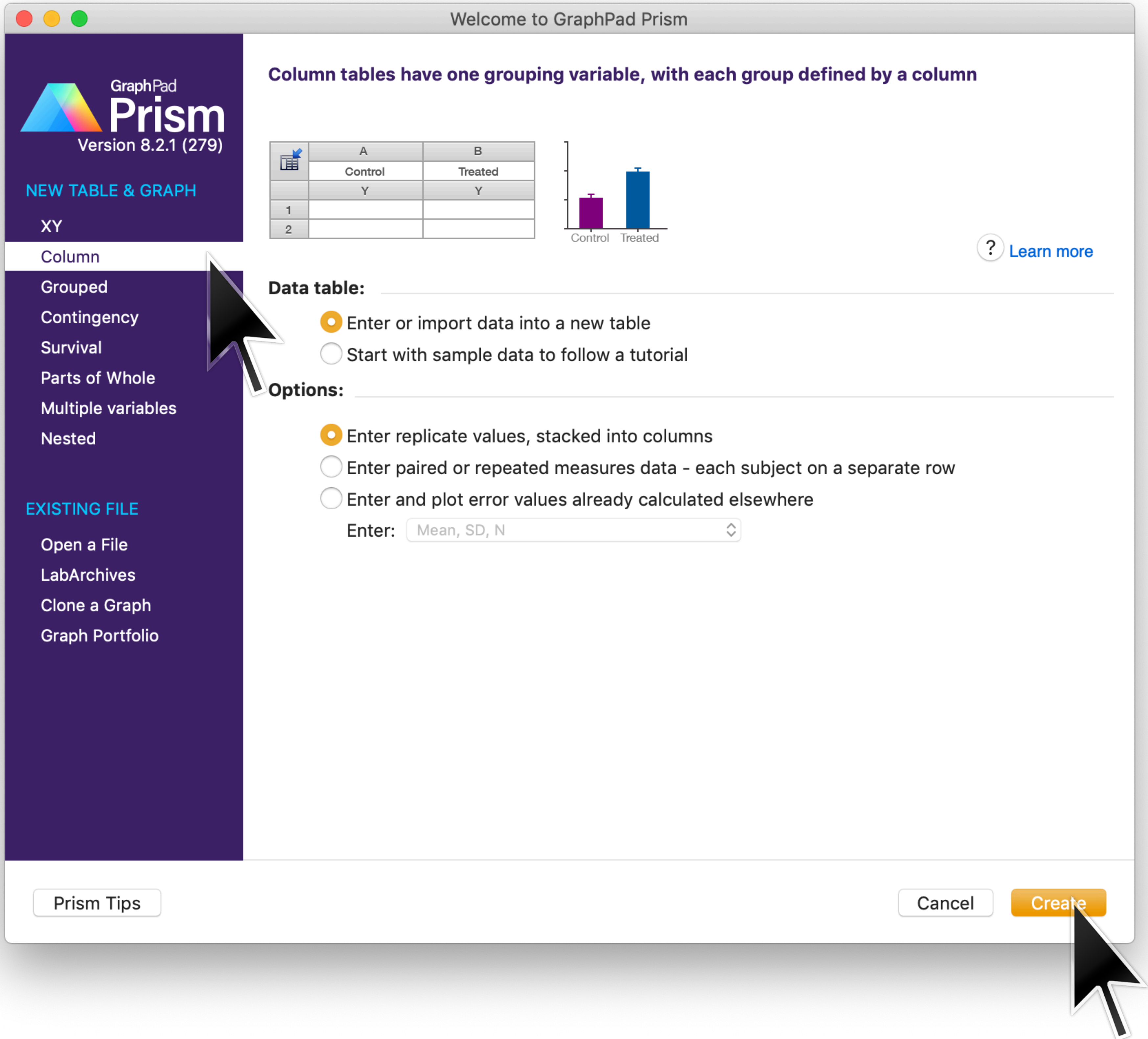
95% CI: [98.12 F, 98.38 F]

temperature standard deviation σ

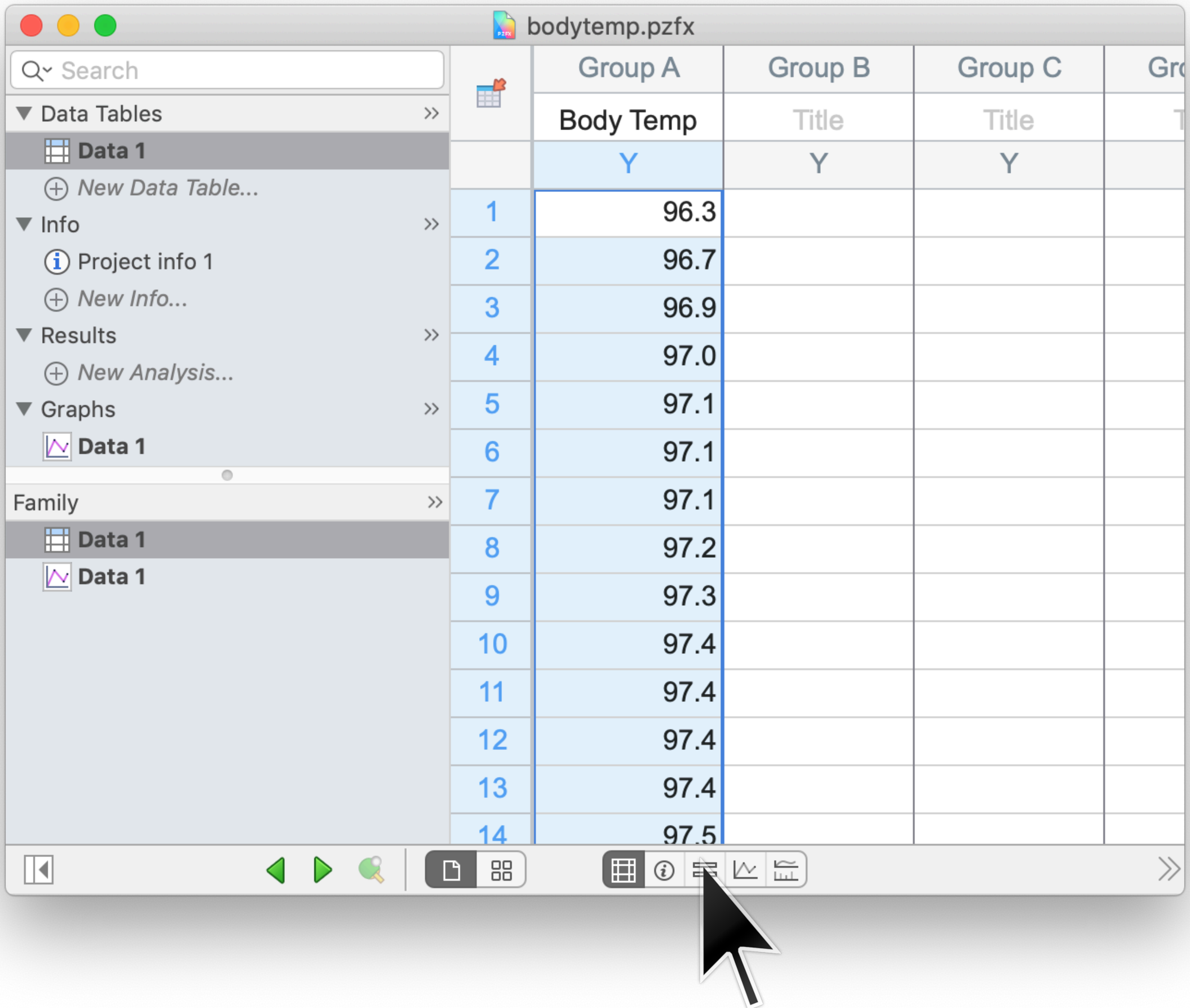
estimate: 0.73 F

95% CI: [0.65 F, 0.83 F]

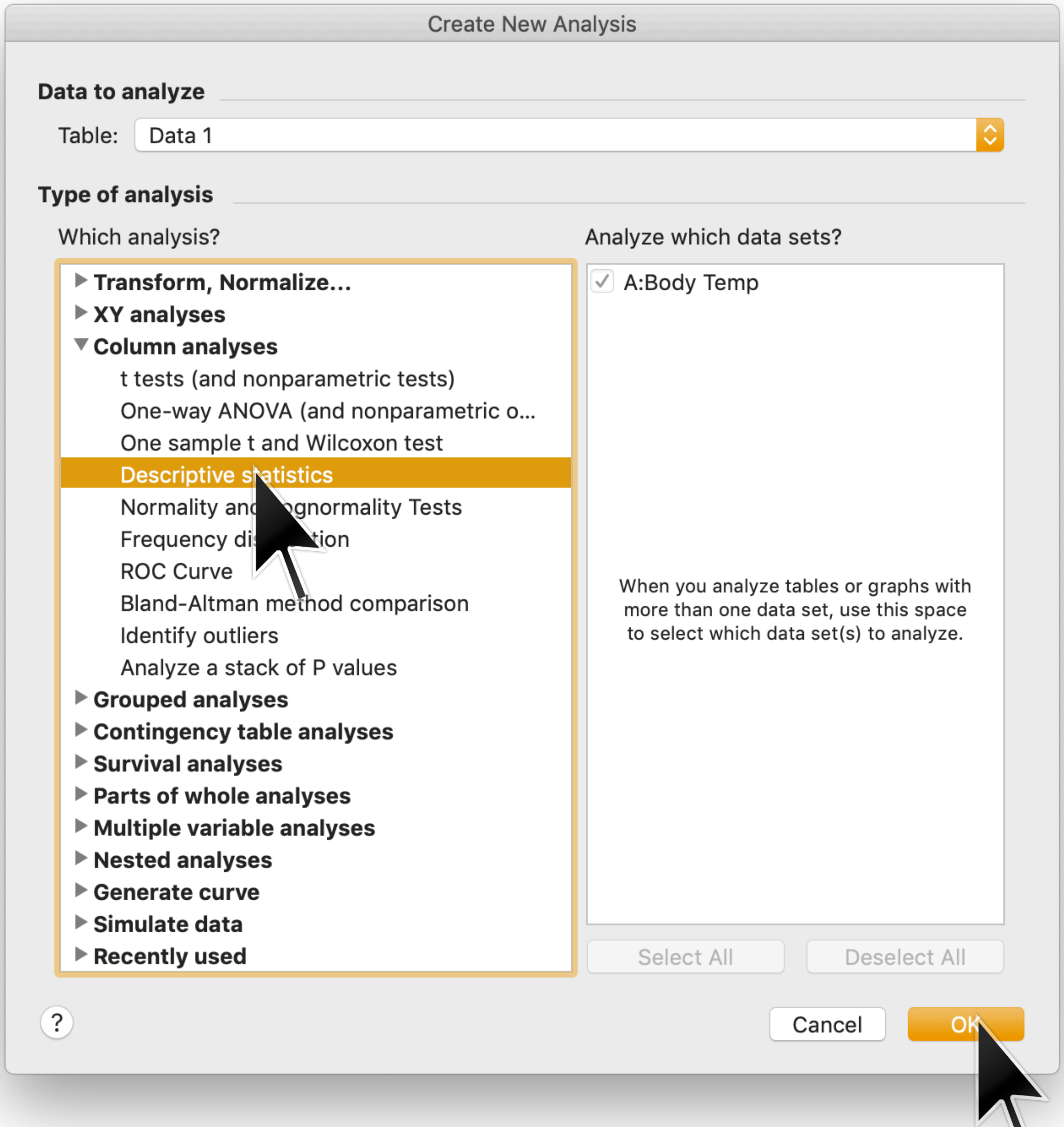
How to do this in PRISM



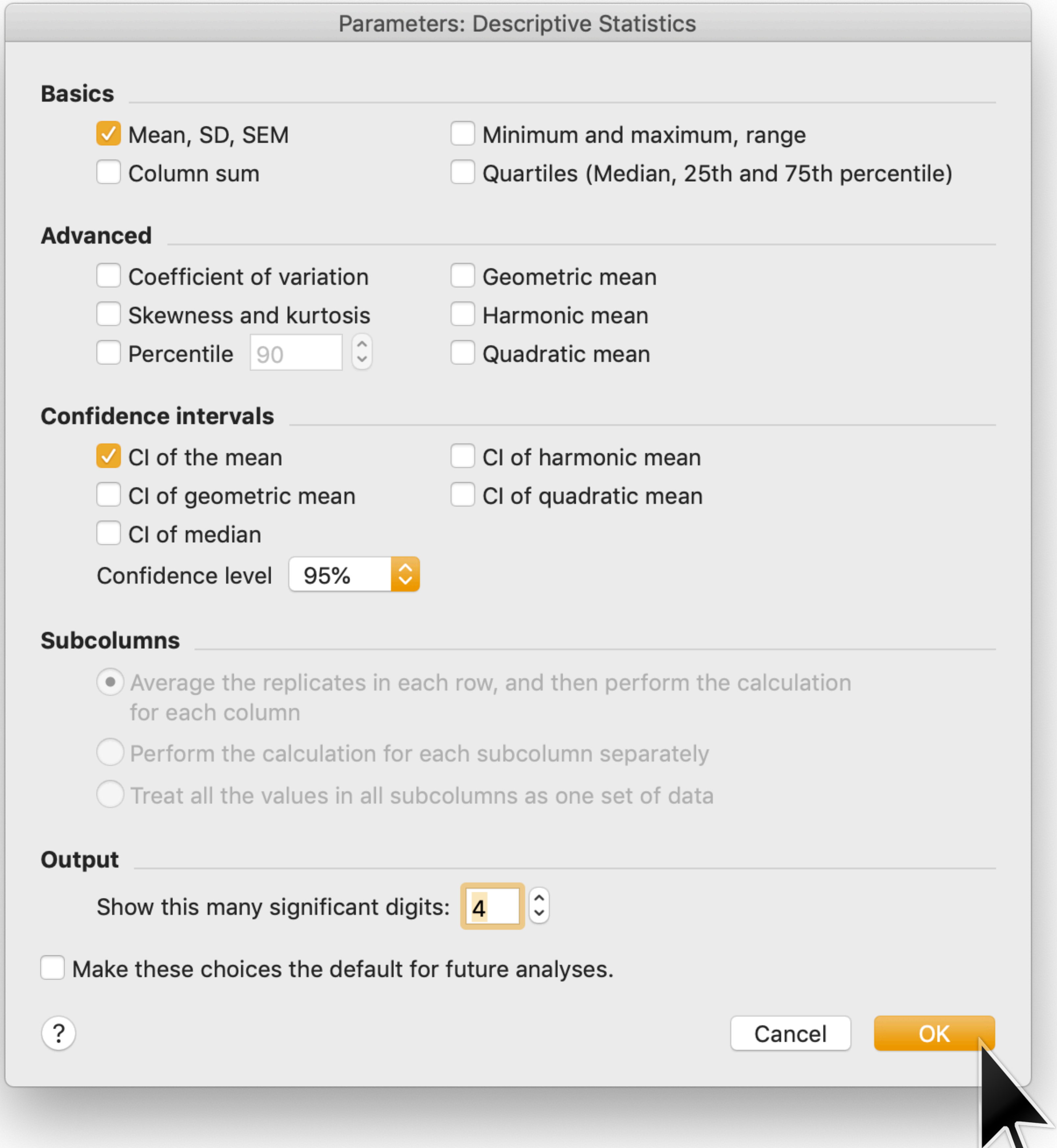
How to do this in PRISM



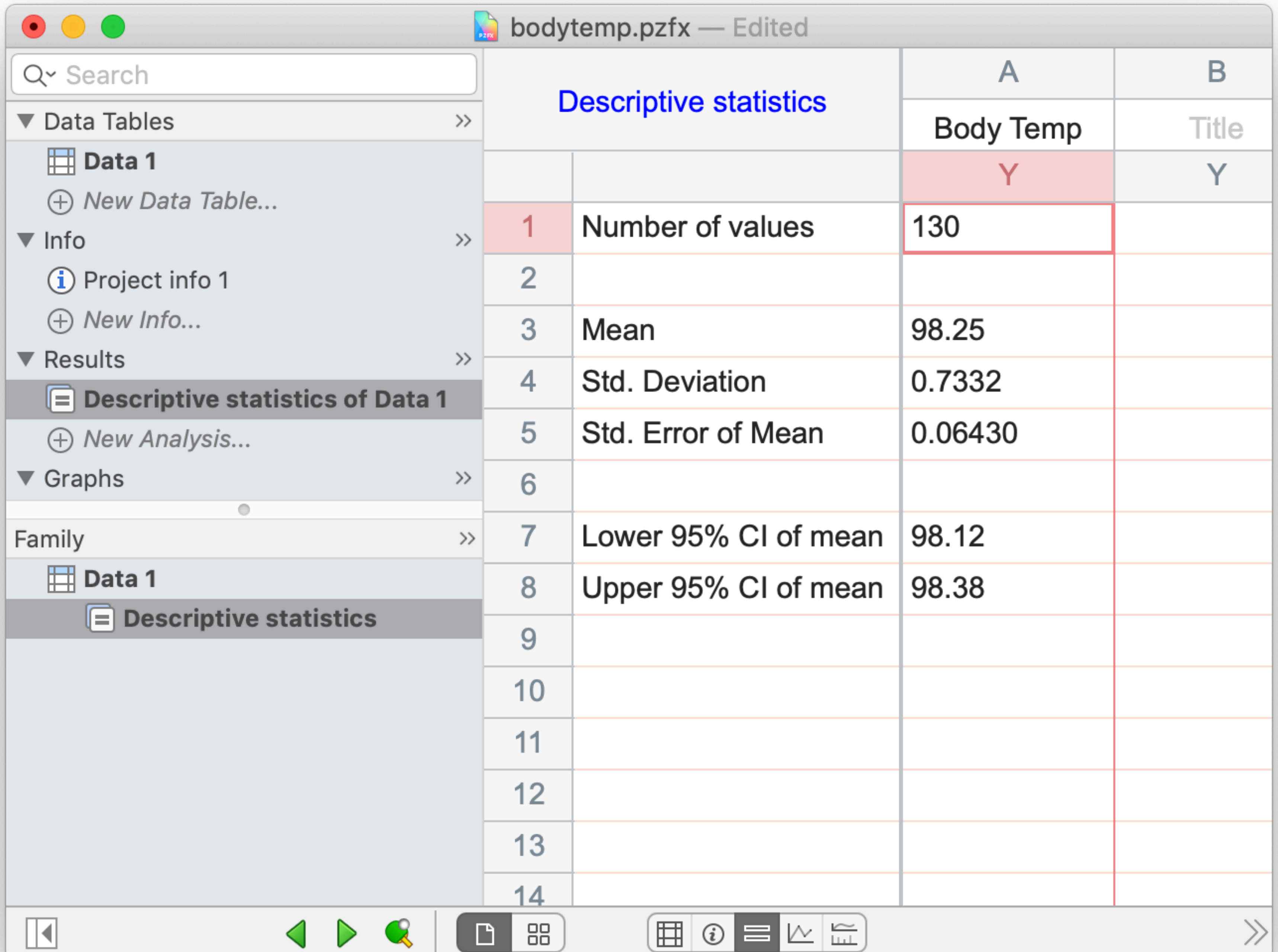
How to do this in PRISM



How to do this in PRISM



How to do this in PRISM



We assume the temperature of a healthy person is drawn from a normal distribution

data

$$x_1, x_2, \ldots, x_N$$

x_i: temperature of individual i in Fahrenheit

model

$$x \sim \mathrm{Normal}(\mu, \sigma^2)$$

μ: average body temperature

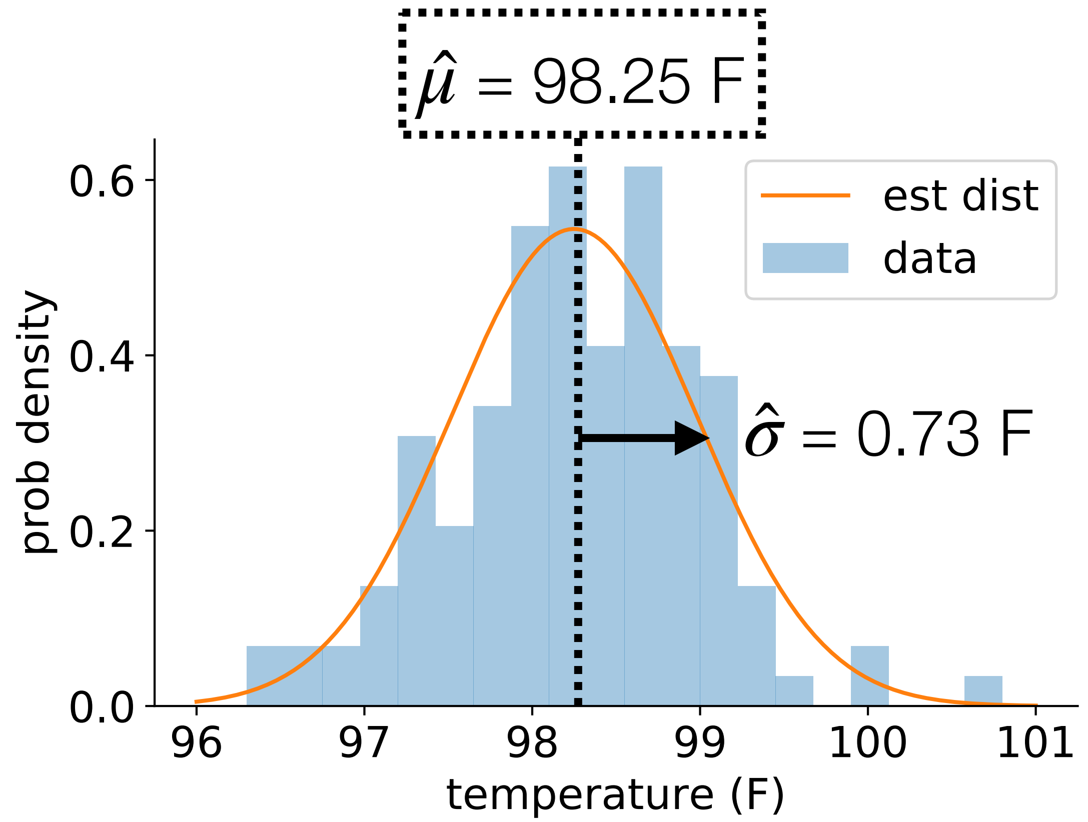
σ: standard deviation of temperatures

goals

1. Compute best estimates for both μ, σ

2. Compute confidence intervals for both μ, σ

We want to infer two parameters from our data



Here there are two parameters that need to be estimated, μ and σ

This is unlike with the binomial distribution, where there was only one parameter q.

The lingering uncertainty in μ is described by a t-distribution

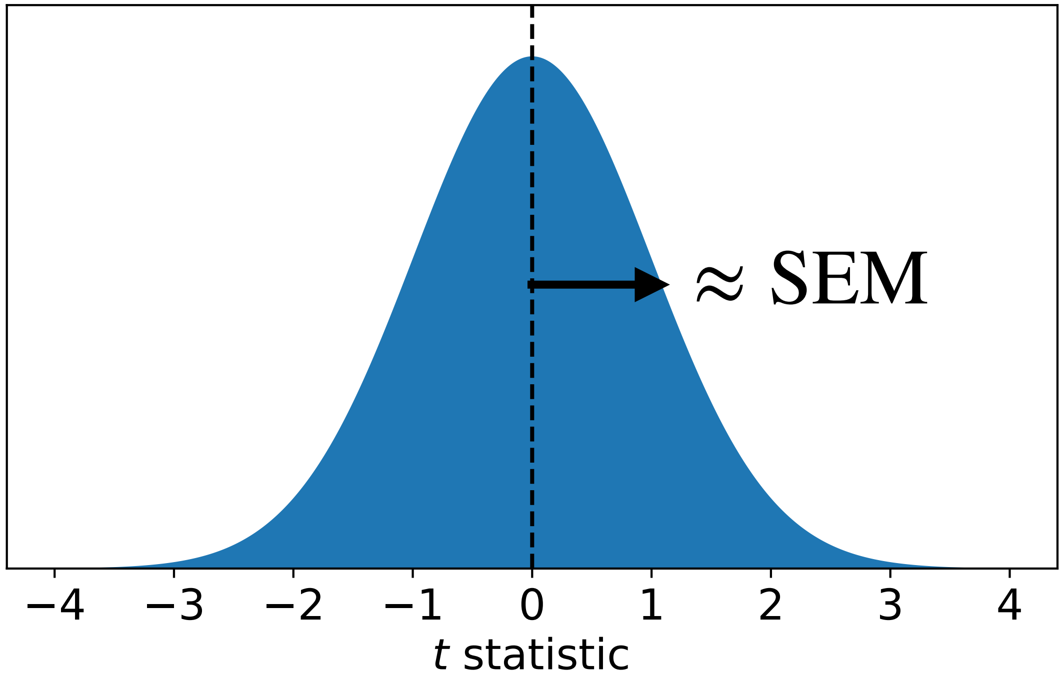
The standard error of the mean (SEM) is given by

$$\text{SEM} = \frac{\hat{\sigma}}{\sqrt{N}}$$

A t-statistic is then used to indicate how strongly μ deviates from $\hat{\mu}$:

$$t = \frac{\mu - \hat{\mu}}{\text{SEM}}$$

The t-statistic follows a t-distribution
(almost a normal distribution, but not quite)

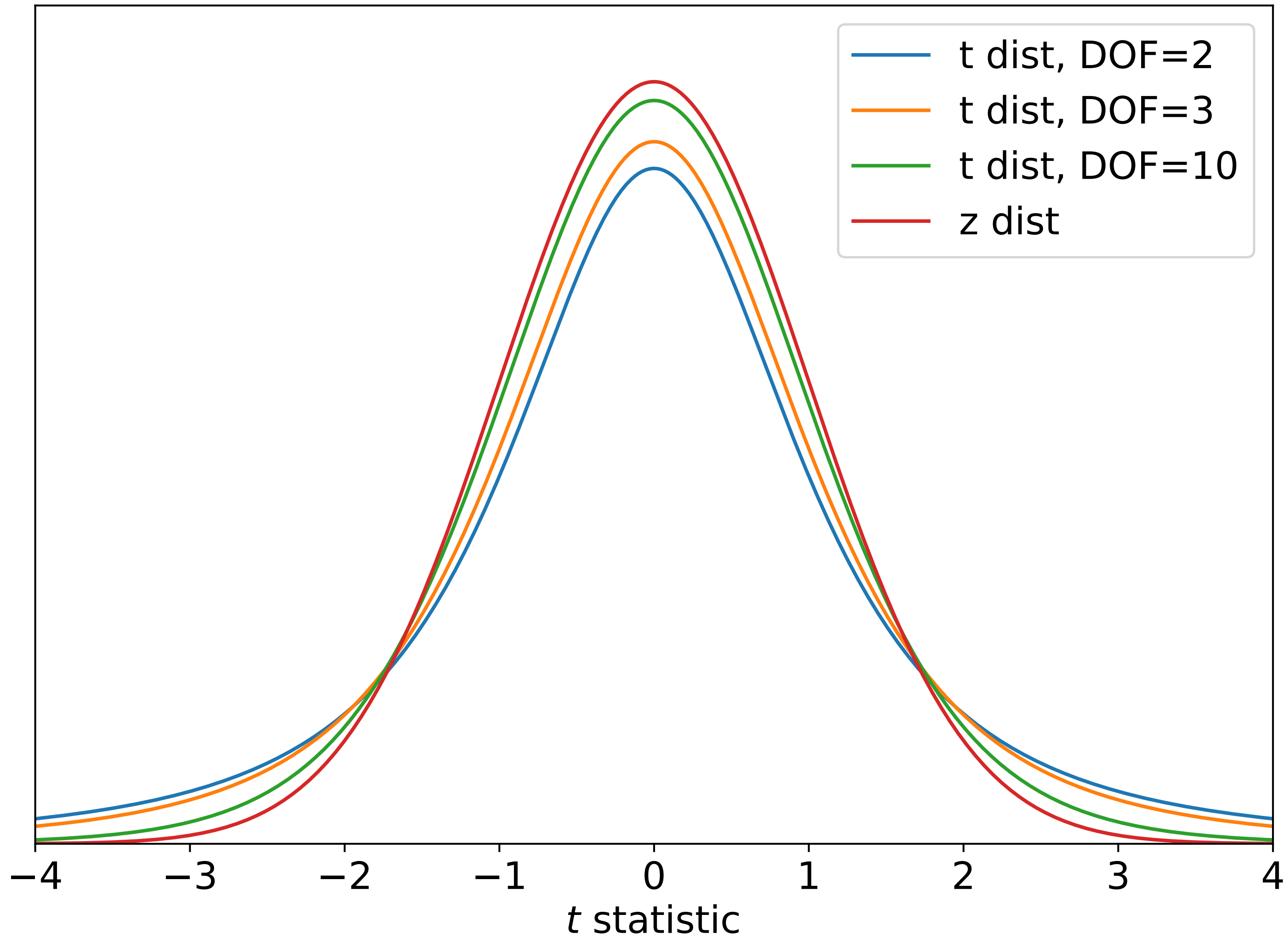


The shape of the t-distribution is affected by the number of degrees of freedom (DOF)

In this case, we use a t-distribution with DOF given by

$$\mathrm{DOF} = N - 1$$

This is almost indistinguishable from a normal (z) distribution when $\mathrm{DOF} \gtrsim 10$.

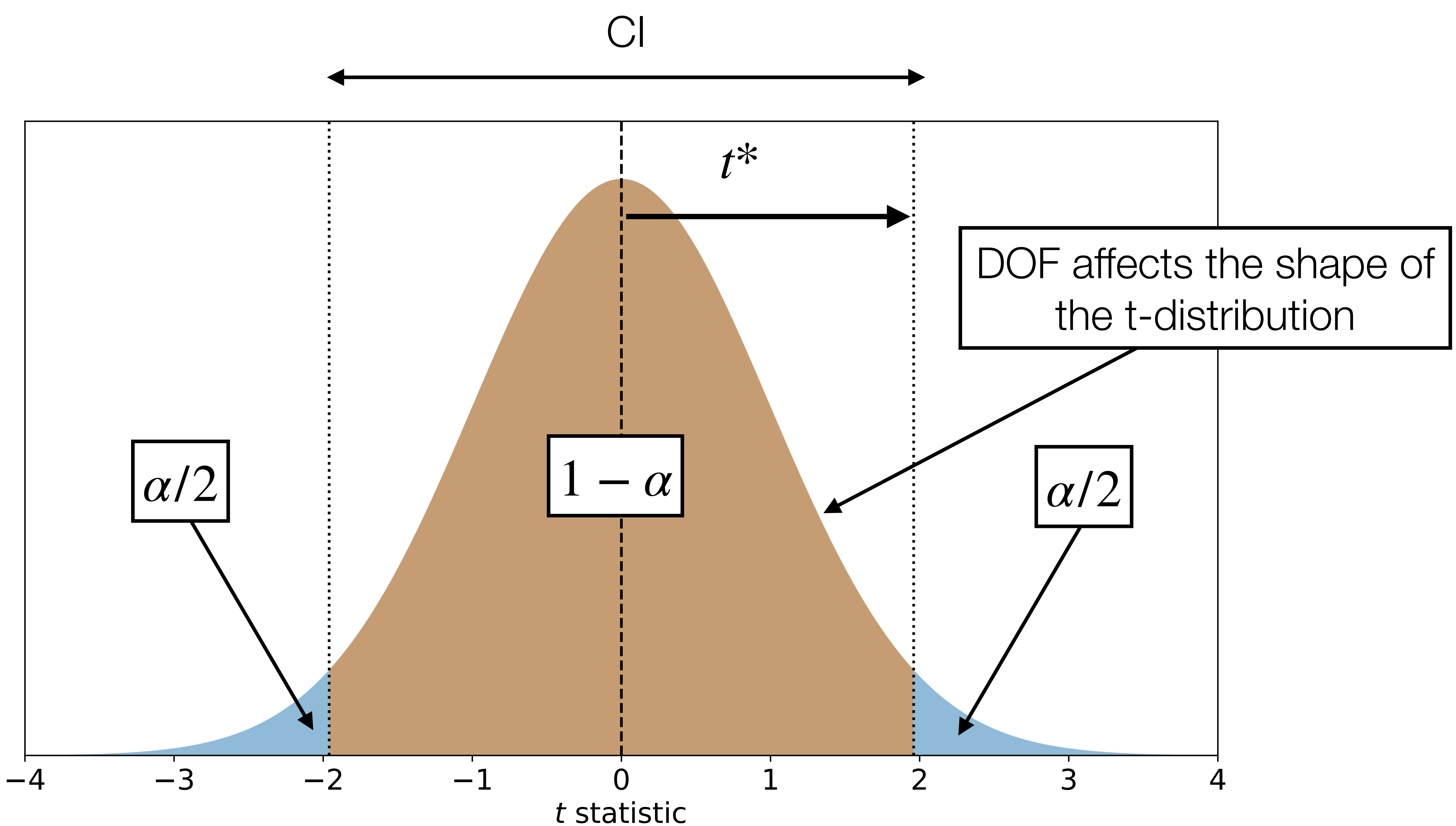


The t-distribution is used to compute a t-statistic cutoff, which determines the confidence interval

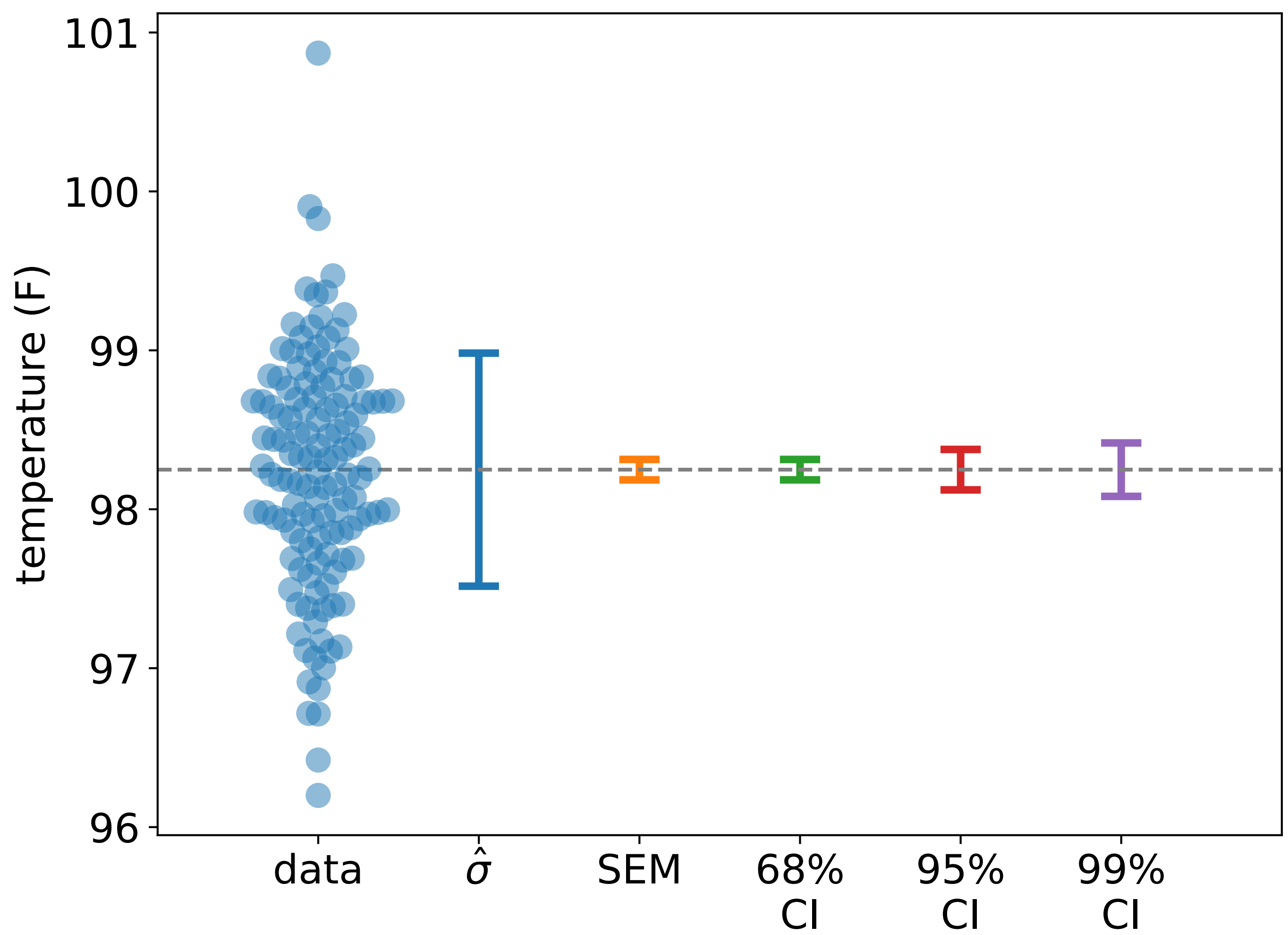
The t-statistic cutoff, t^*, is determined by both α and the DOF.

margin of error: $W = t^* \cdot \text{SEM}$

confidence interval: $\hat{\mu} \pm W$



Confidence intervals (CIs) and standard errors of the mean (SEMs) quantify how uncertain a parameter



SEMs and CIs of the mean quantify the uncertainty in μ, not the width of the sampling distribution ($\hat{\sigma}$).

SEMs and CIs <u>decrease</u> in size as the amount of data <u>increases</u>.

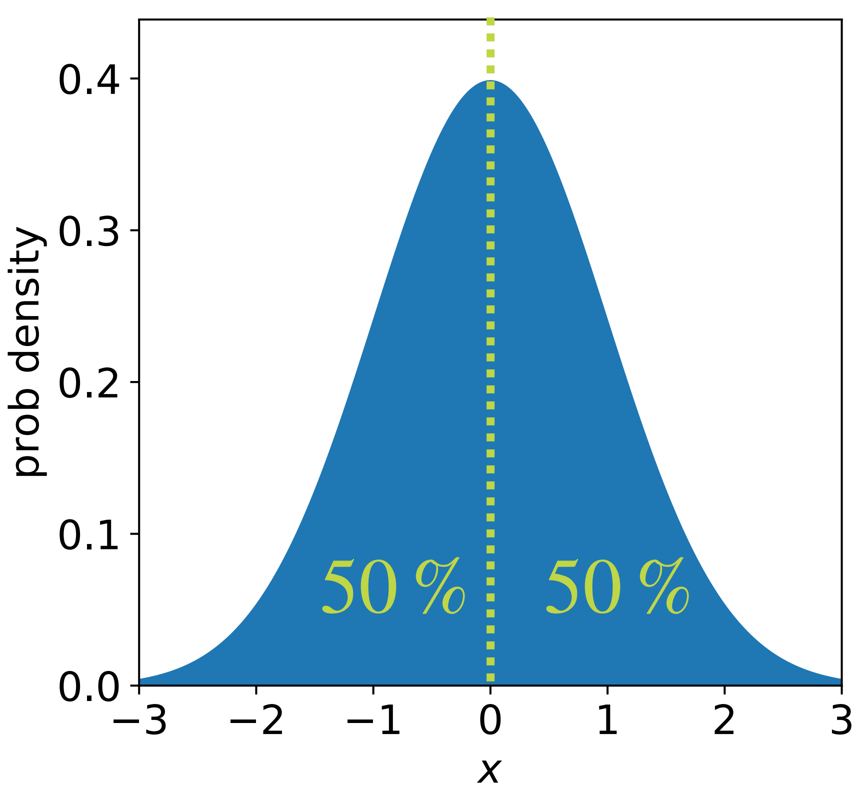
CIs <u>increase</u> in size if the required confidence level <u>increases</u> (i.e., α decreases)

The median is the standard nonparametric estimate of a distribution's center

For data: sort the data $x_1, x_2, x_3, \ldots, x_N$ in ascending order. The median is then defined as:

$$\text{median} = q_{50} = \begin{cases} x_{\frac{N+1}{2}} & \text{if} \quad N \text{ odd} \\ \frac{1}{2}\left(x_{\frac{N}{2}} + x_{\frac{N+2}{2}}\right) & \text{if} \quad N \text{ even} \end{cases}$$

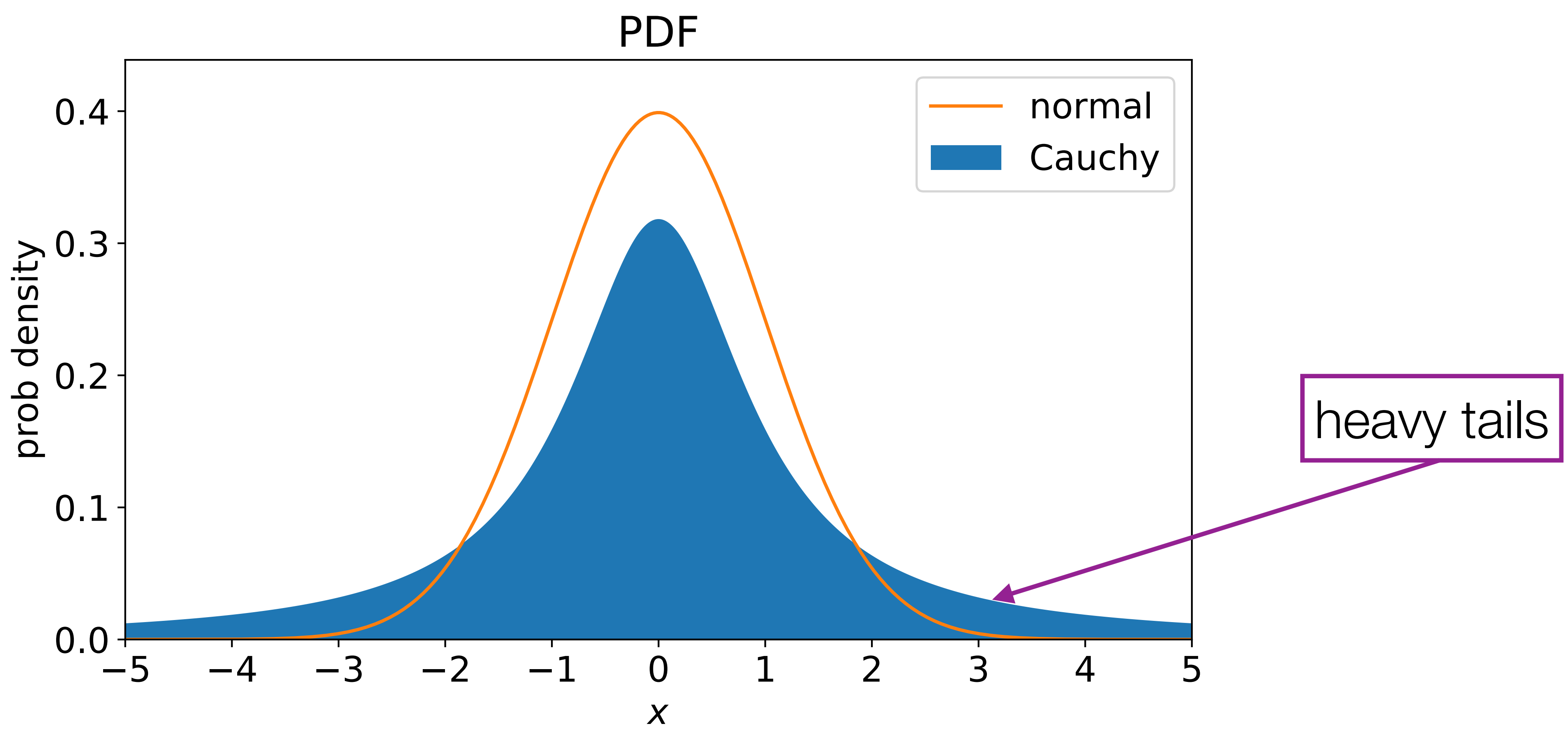
For a distribution: the median is the value of x that separates half the distribution's mass from the other.



The median of a symmetric distribution is equal to its mean

The median is less sensitive to outliers than the mean

Cauchy distribution



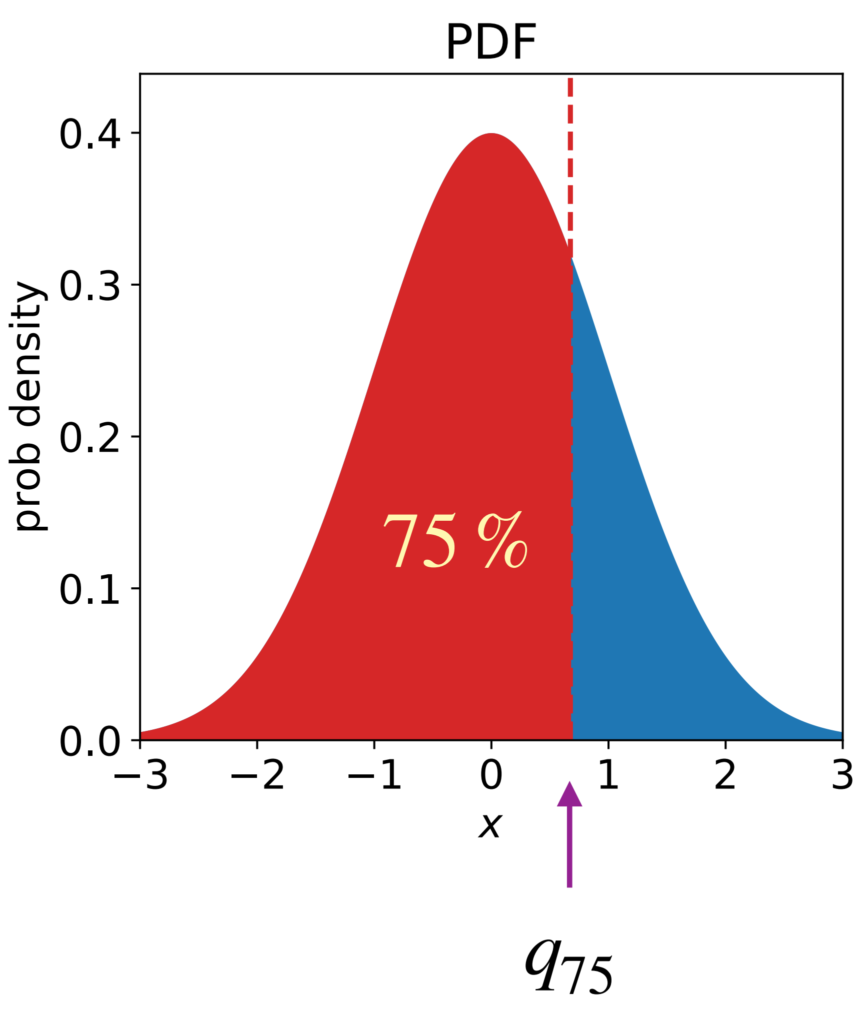
The standard estimate of the mean $\hat{\mu}$ will <u>not converge</u> as N becomes large!

The median q_{50} <u>does converge,</u> just as quickly as for any distribution.

Quantiles of a distribution

More generally, the quantile q_K of a distribution is the value of x that bounds $K\,\%$ of the distribution's mass.

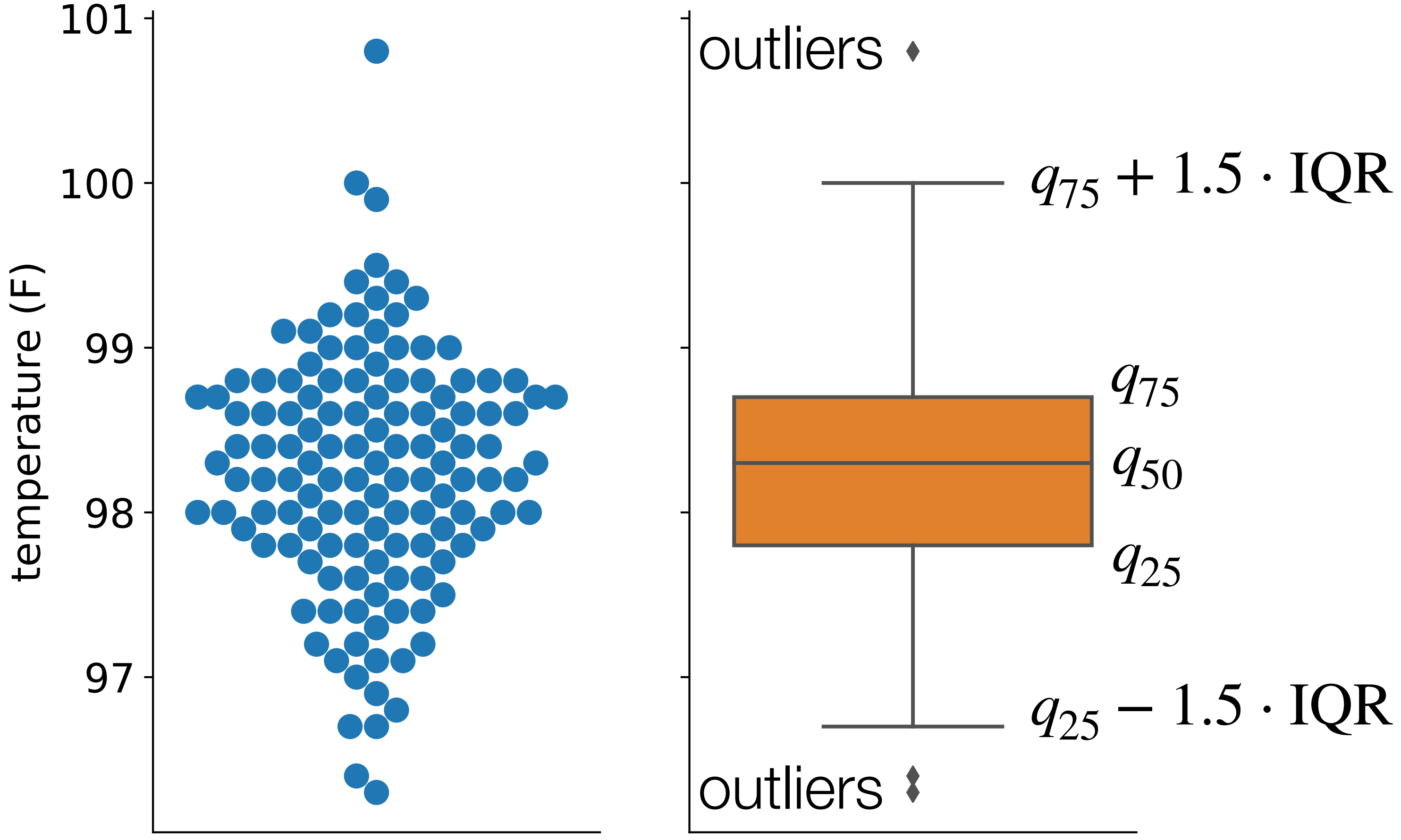
E.g., the median in the quantile q_{50}



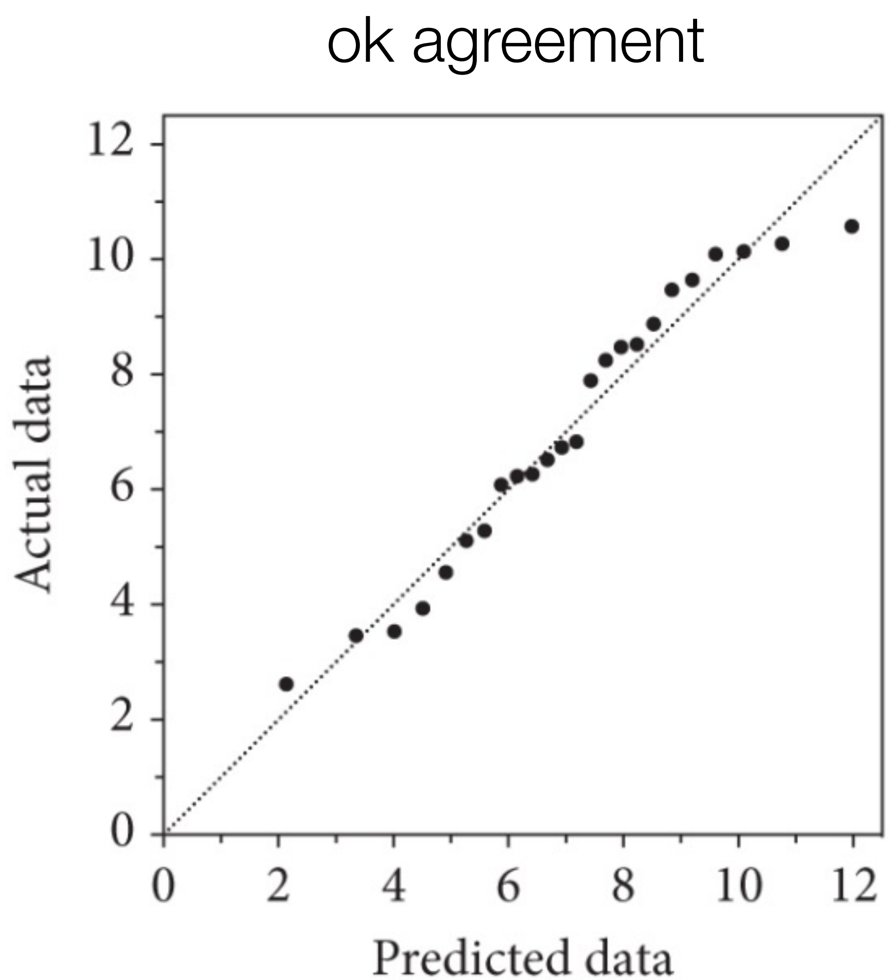
Box and whisker plots indicate quantiles

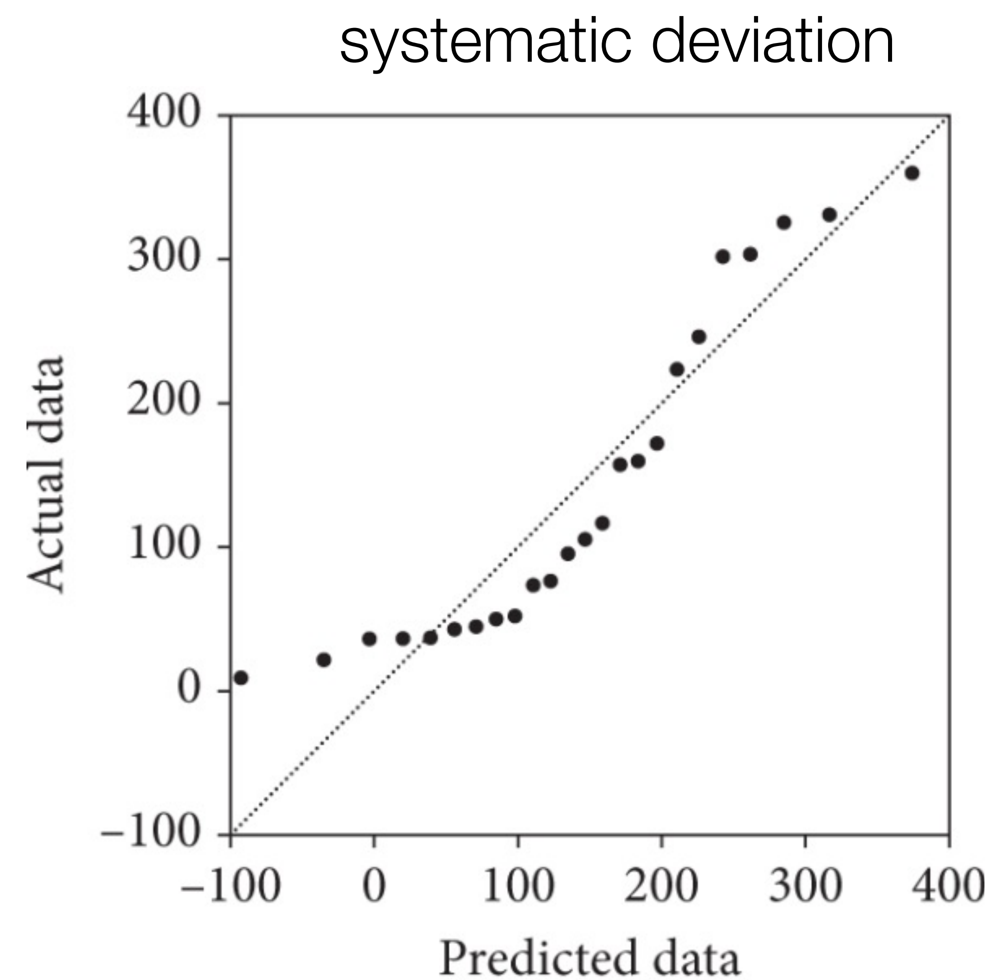
Interquartile range is defined by

$$\mathrm{IQR} = q_{75} - q_{25}$$



QQ plots are used to visually test whether data follows an expected distribution.



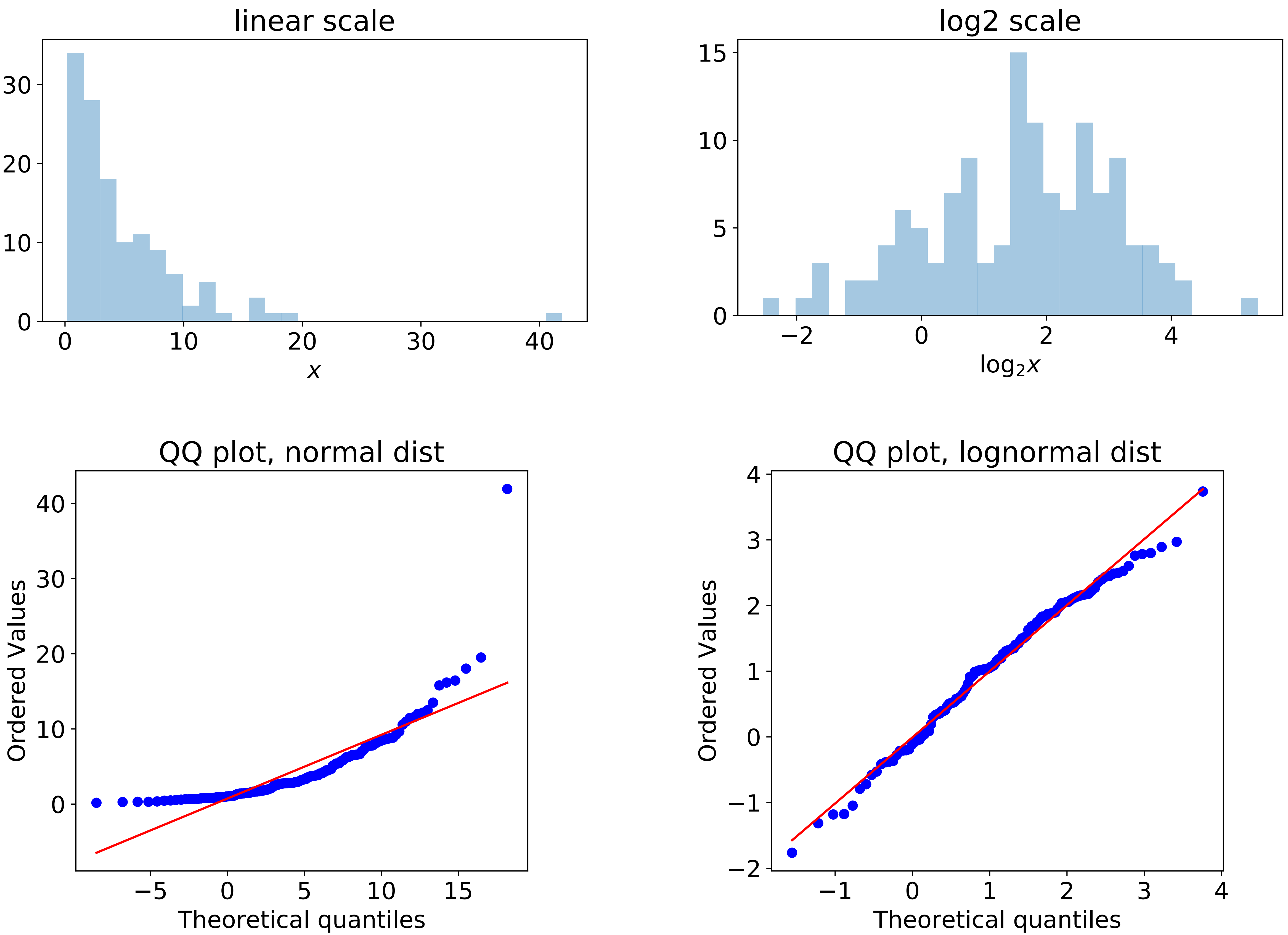


y axis: sorted data values $x_1, x_2, \ldots, N$.

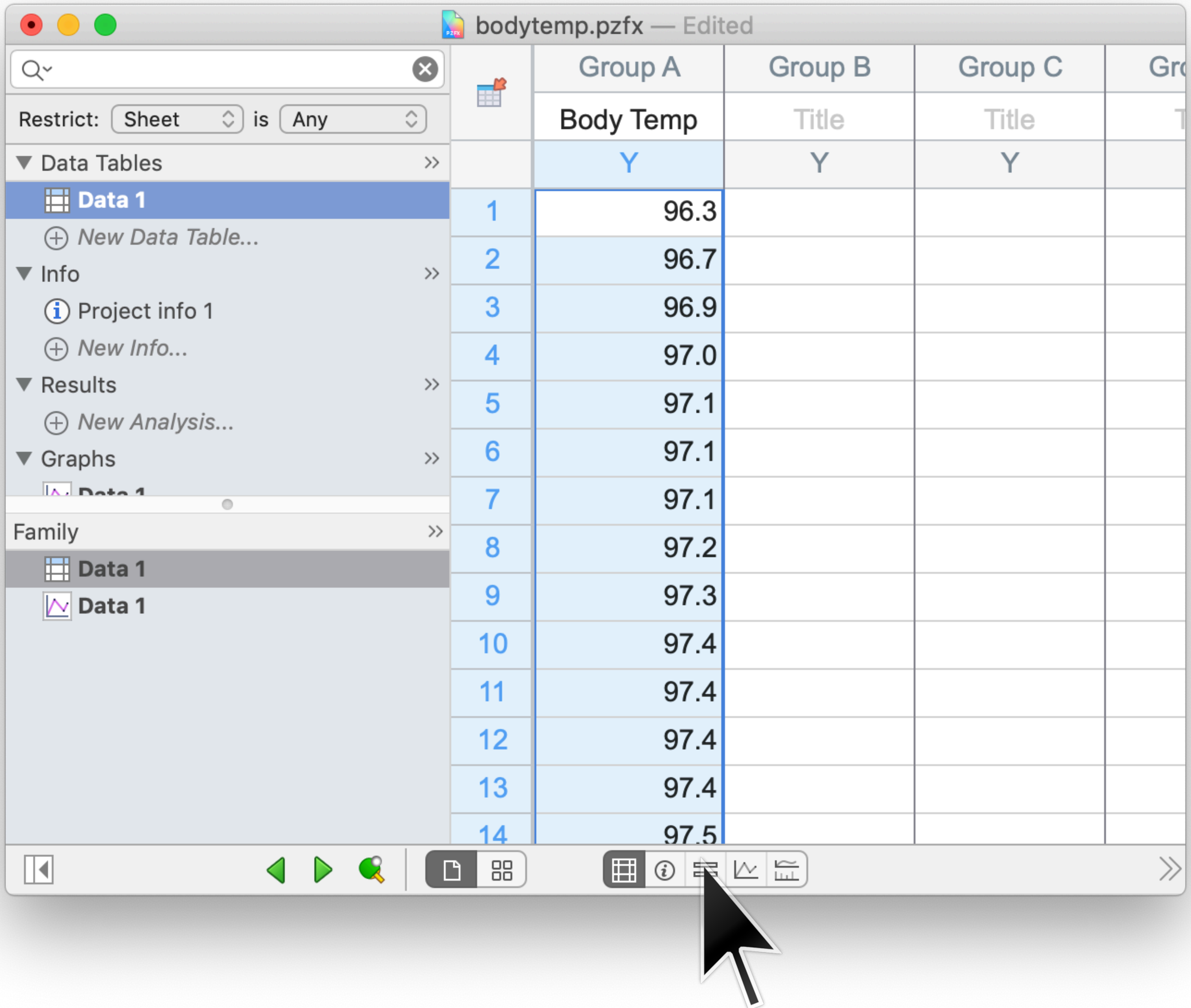
x axis: corresponding quantiles q_X of the inferred distribution, using the percentile values $X_1, X_2, \ldots, X_N$ computed for each data point.

The analysis of a QQ plot is done by eye and making a judgement call.

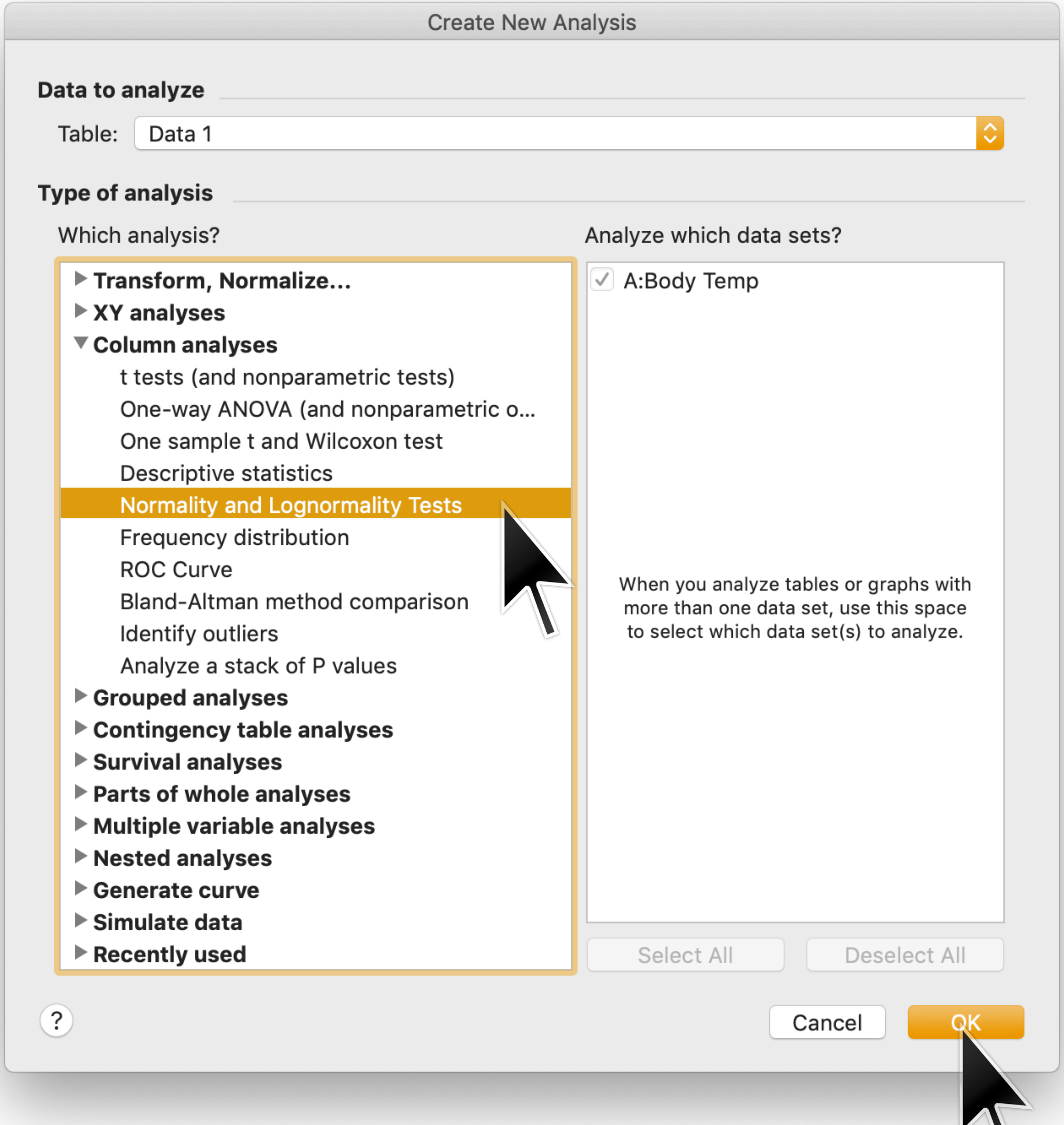
QQ plot example: simulated lognormal data



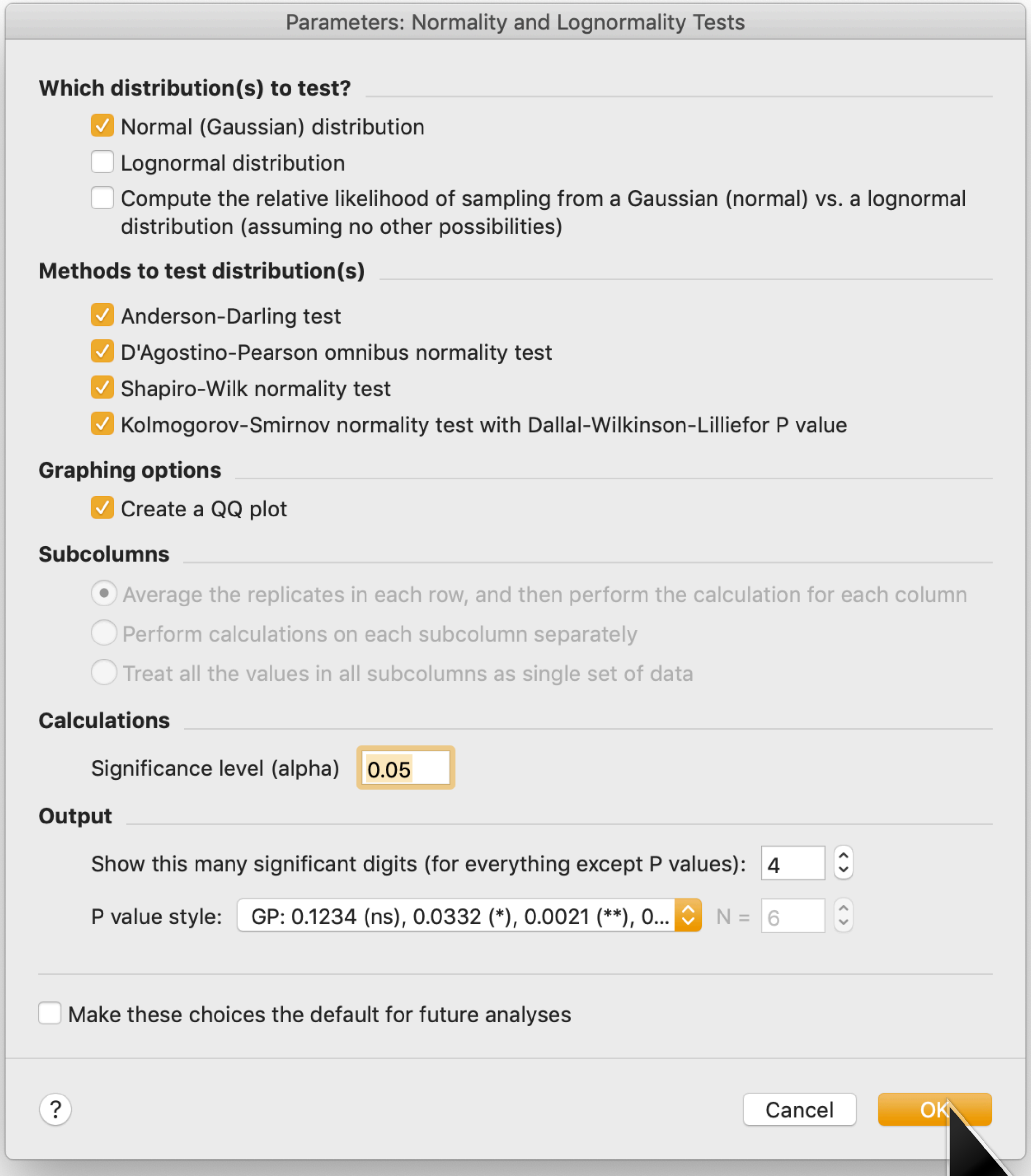
How to do this in Prism



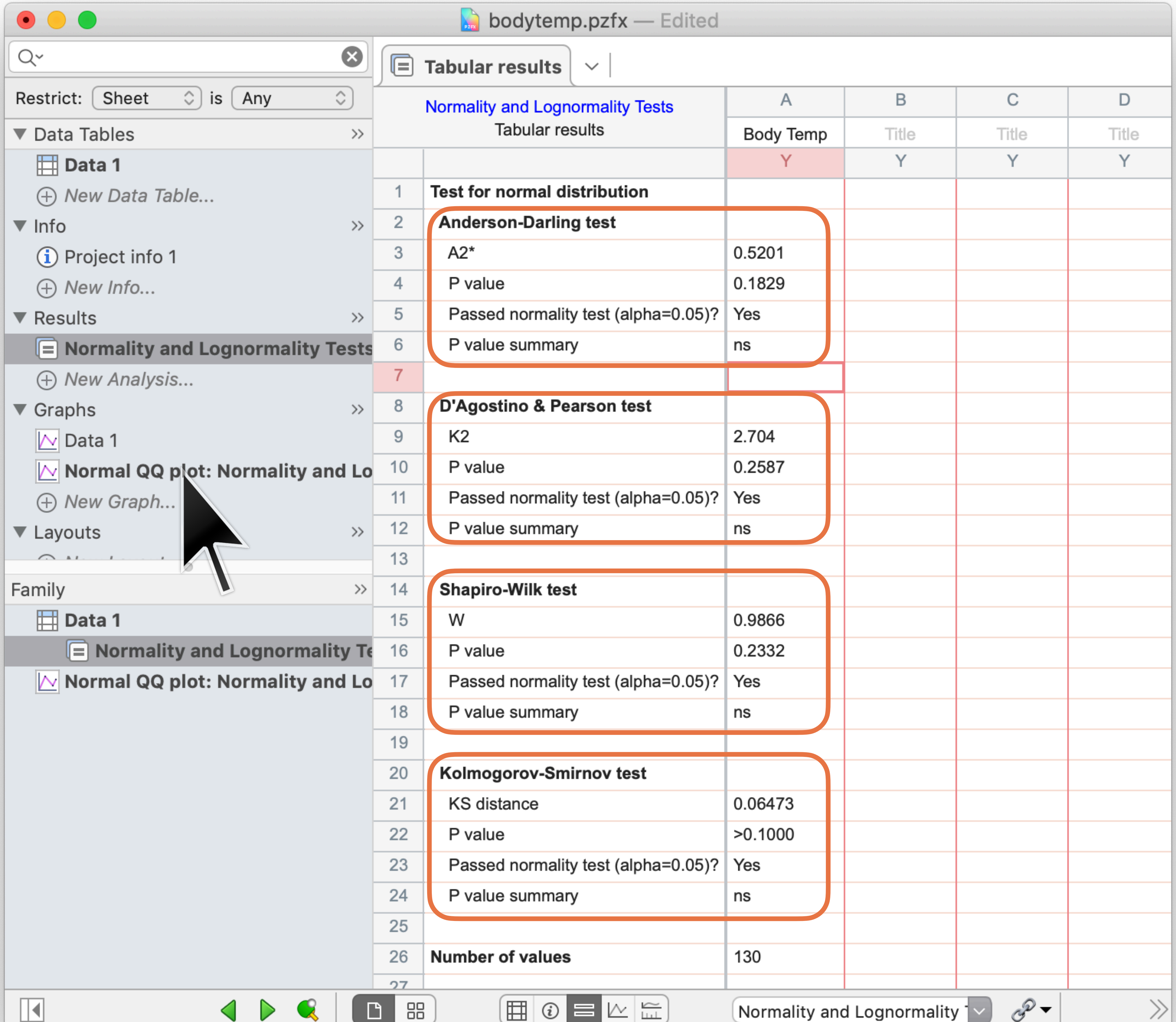
How to do this in Prism



How to do this in Prism

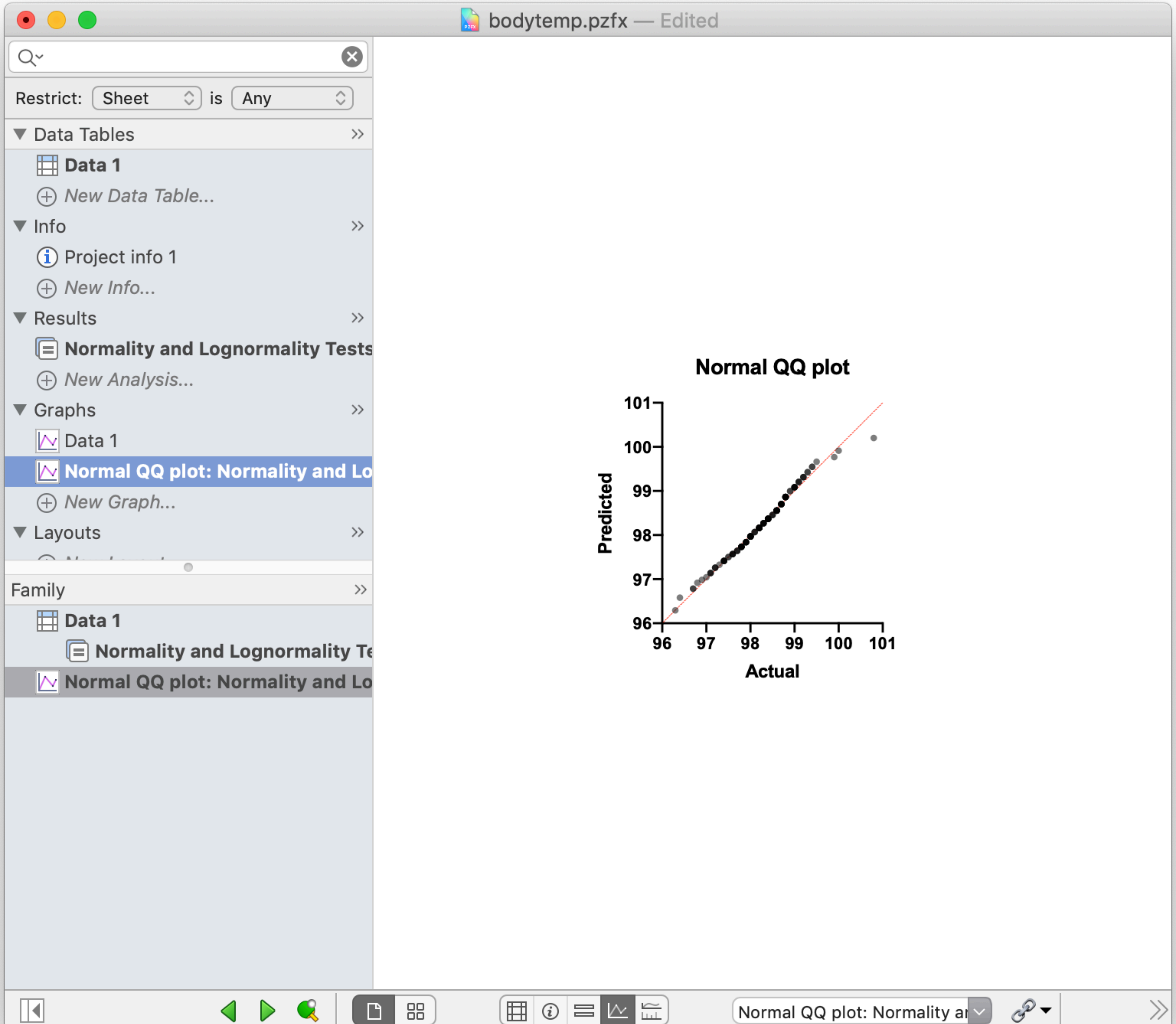


How to do this in Prism

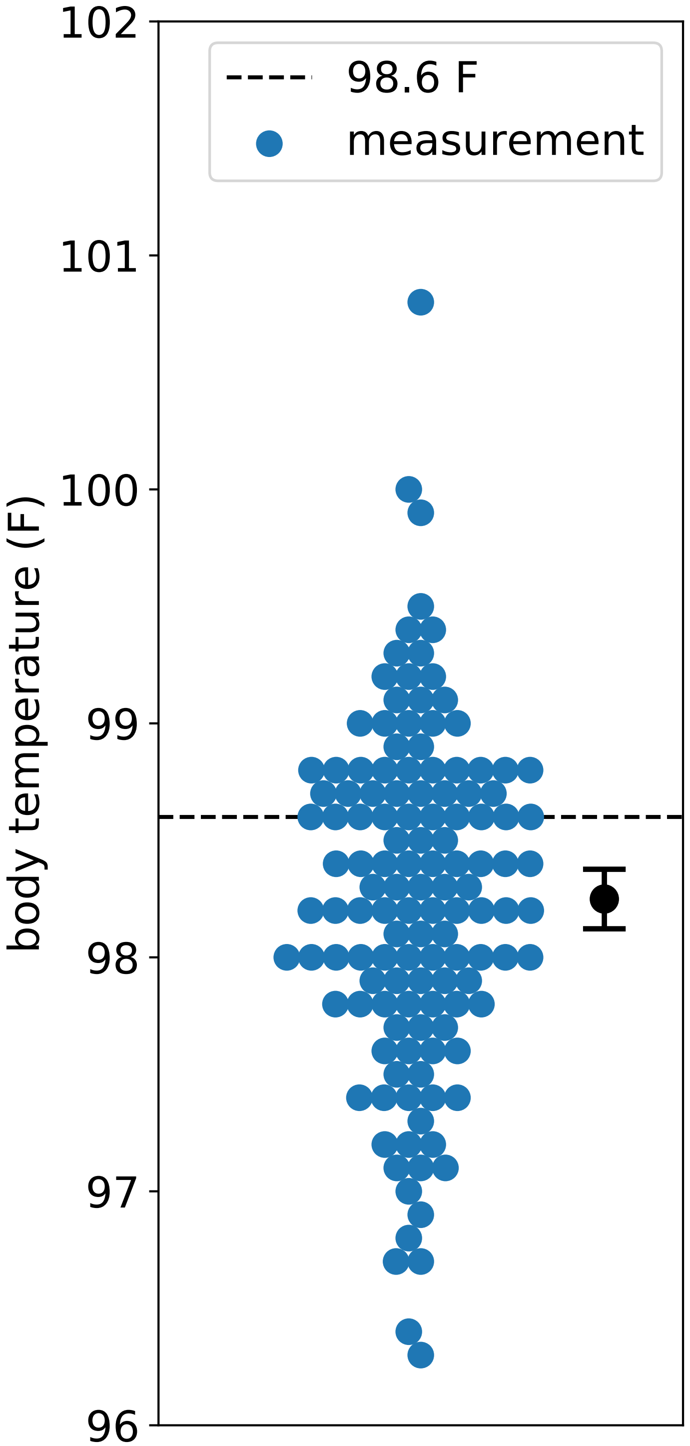


	Normality and Lognormality Tests Tabular results	A Body Temp Y	B Title Y	C Title Y	D Title Y
1	**Test for normal distribution**				
2	**Anderson-Darling test**				
3	A2*	0.5201			
4	P value	0.1829			
5	Passed normality test (alpha=0.05)?	Yes			
6	P value summary	ns			
7					
8	**D'Agostino & Pearson test**				
9	K2	2.704			
10	P value	0.2587			
11	Passed normality test (alpha=0.05)?	Yes			
12	P value summary	ns			
13					
14	**Shapiro-Wilk test**				
15	W	0.9866			
16	P value	0.2332			
17	Passed normality test (alpha=0.05)?	Yes			
18	P value summary	ns			
19					
20	**Kolmogorov-Smirnov test**				
21	KS distance	0.06473			
22	P value	>0.1000			
23	Passed normality test (alpha=0.05)?	Yes			
24	P value summary	ns			
25					
26	**Number of values**	130			

How to do this in PRISM



Student's t test (one sample)



Null Hypothesis:

a population is normally distributed with a known mean value of μ_{null}

Data:

measurements: $x_1, x_2, \ldots, x_N$

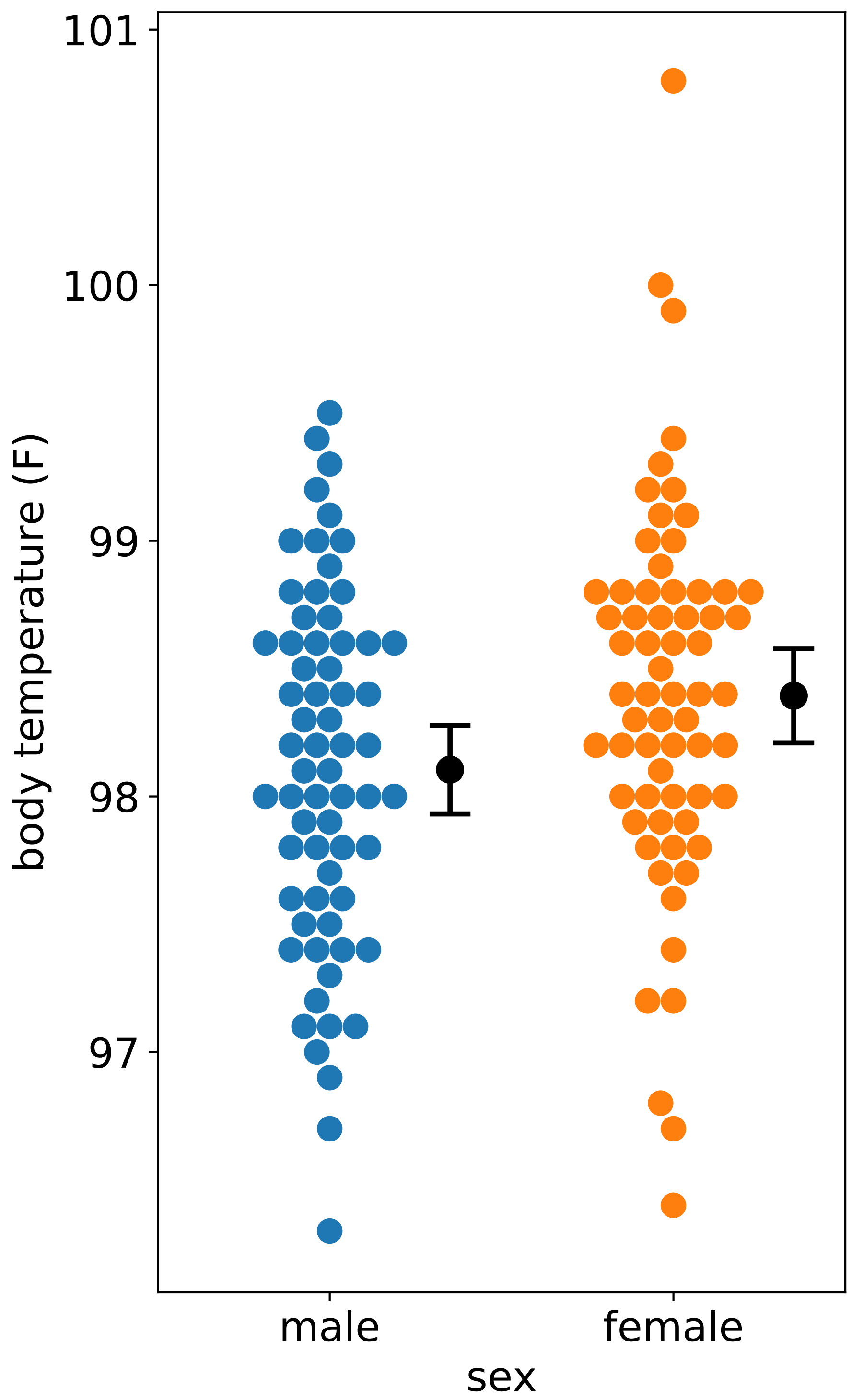
Test statistic:

$$t = \frac{\hat{\mu} - \mu_{\text{null}}}{\text{SEM}}$$

Null distribution:

t distribution with $\text{DOF} = N - 1$.

Student's t test (two sample, equal variance)



Null Hypothesis:

two populations have the same mean

Data:

$x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$

Assumptions:

the two populations follow normal distributions and have equal variances

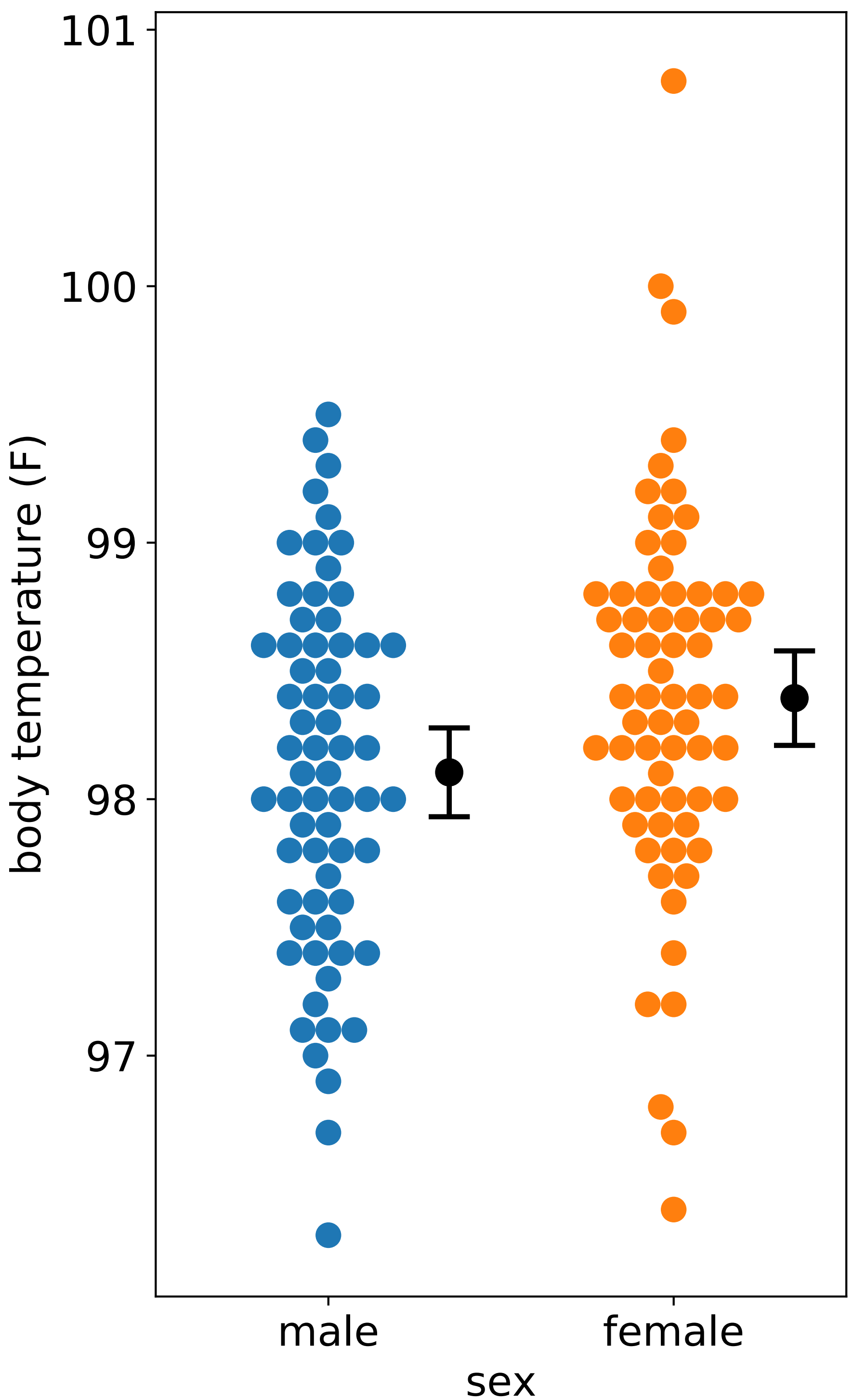
Test statistic:

$$t = \frac{\hat{\mu}_x - \hat{\mu}_y}{\hat{\sigma}\sqrt{\frac{1}{m} + \frac{1}{n}}}, \quad \hat{\sigma} = \sqrt{\frac{(m-1)\hat{\sigma}_x^2 + (n-1)\hat{\sigma}_y^2}{m+n-2}}$$

Null distribution:

t distribution with $\mathrm{DOF} = m + n - 2$.

Welch's t test



Null Hypothesis:

two populations have the same mean but
not necessarily the same standard deviation

Data:

$x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$

Advantage:

Fewer assumptions than standard unpaired t test

Disadvantage:

Less power than standard unpaired t tests

Test statistic:

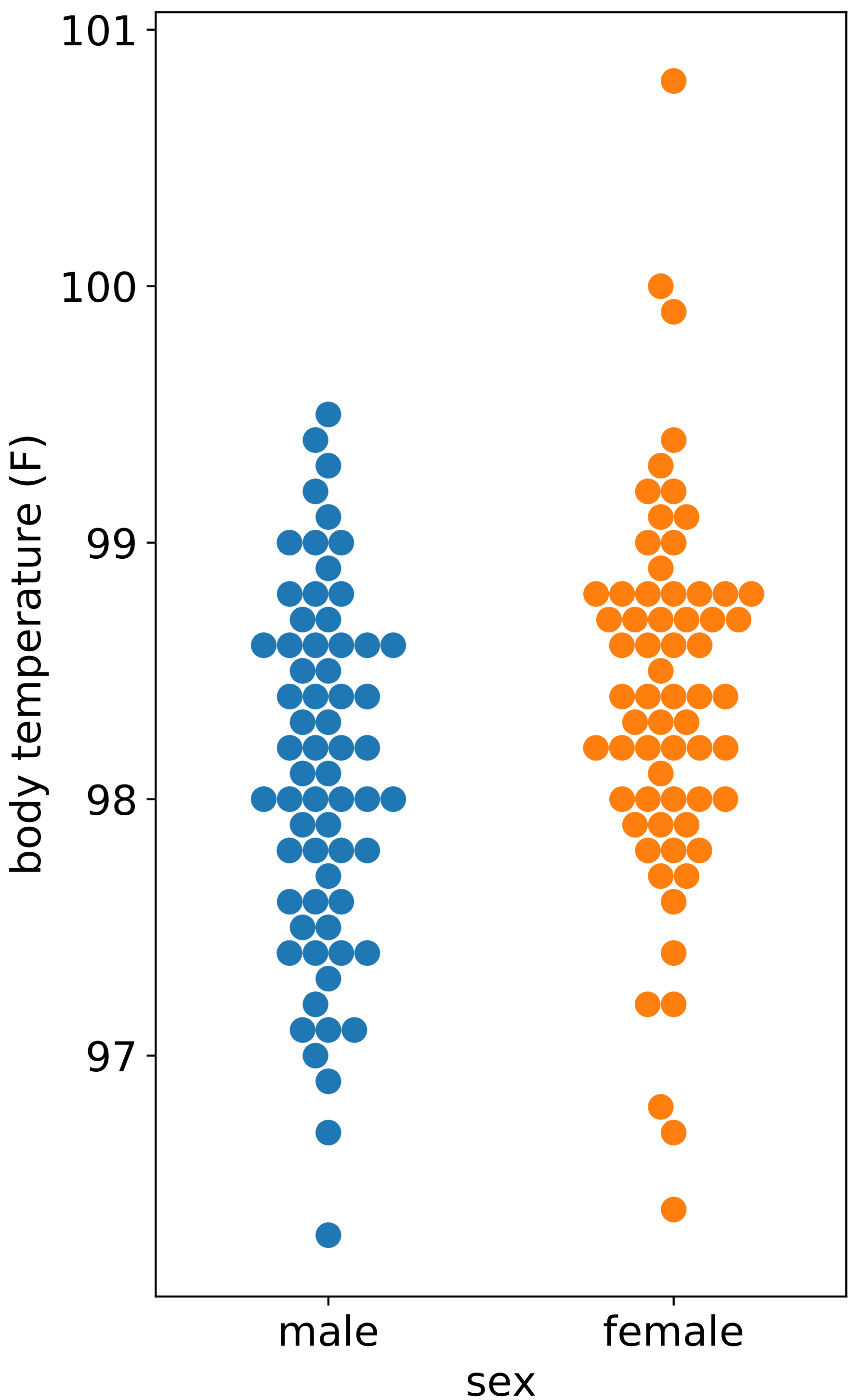
$$t = \frac{\hat{\mu}_x - \hat{\mu}_y}{\sqrt{\frac{\hat{\sigma}_x^2}{m} + \frac{\hat{\sigma}_y^2}{n}}}$$

Null distribution:

Student's t distribution with

$$\text{DOF} = \frac{\left(\frac{\hat{\sigma}_x^2}{m} + \frac{\hat{\sigma}_y^2}{n}\right)^2}{\frac{(\hat{\sigma}_x^2/m)^2}{m-1} + \frac{(\hat{\sigma}_y^2/n)^2}{n-1}}$$

Mann Whitney U test (Wilcoxon rank-sum test)



Null Hypothesis:

If x is sampled from population 1 and y is sampled from population 2,

$$p(x > y) = p(x < y)$$

Data:

$x_1, x_2, \ldots, x_m$ and $y_1, y_2, \ldots, y_n$

Advantage:

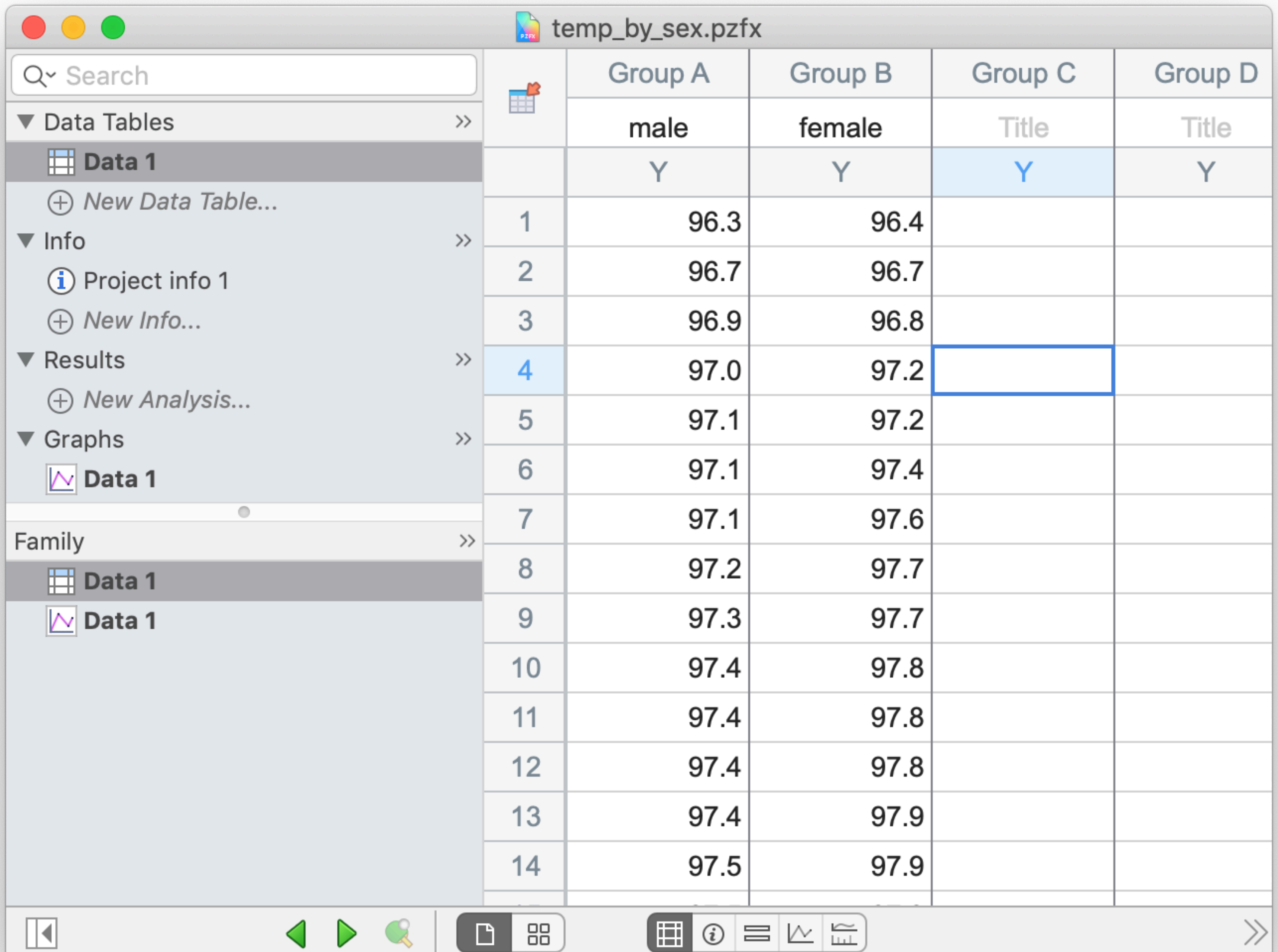
No assumptions about the mathematical form of $p(x)$ and $p(y)$.

Disadvantage:

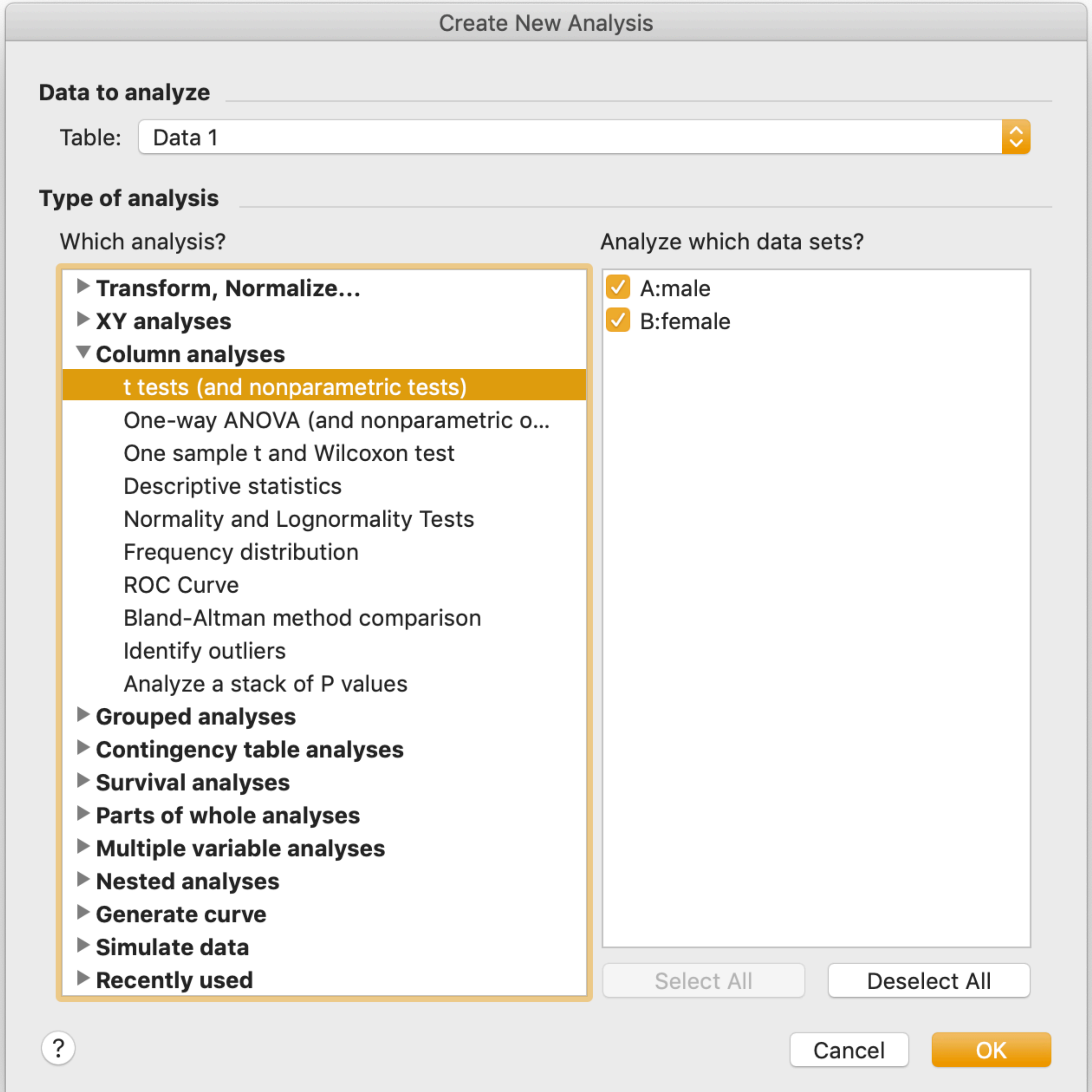
Somewhat less powerful than Student's t test

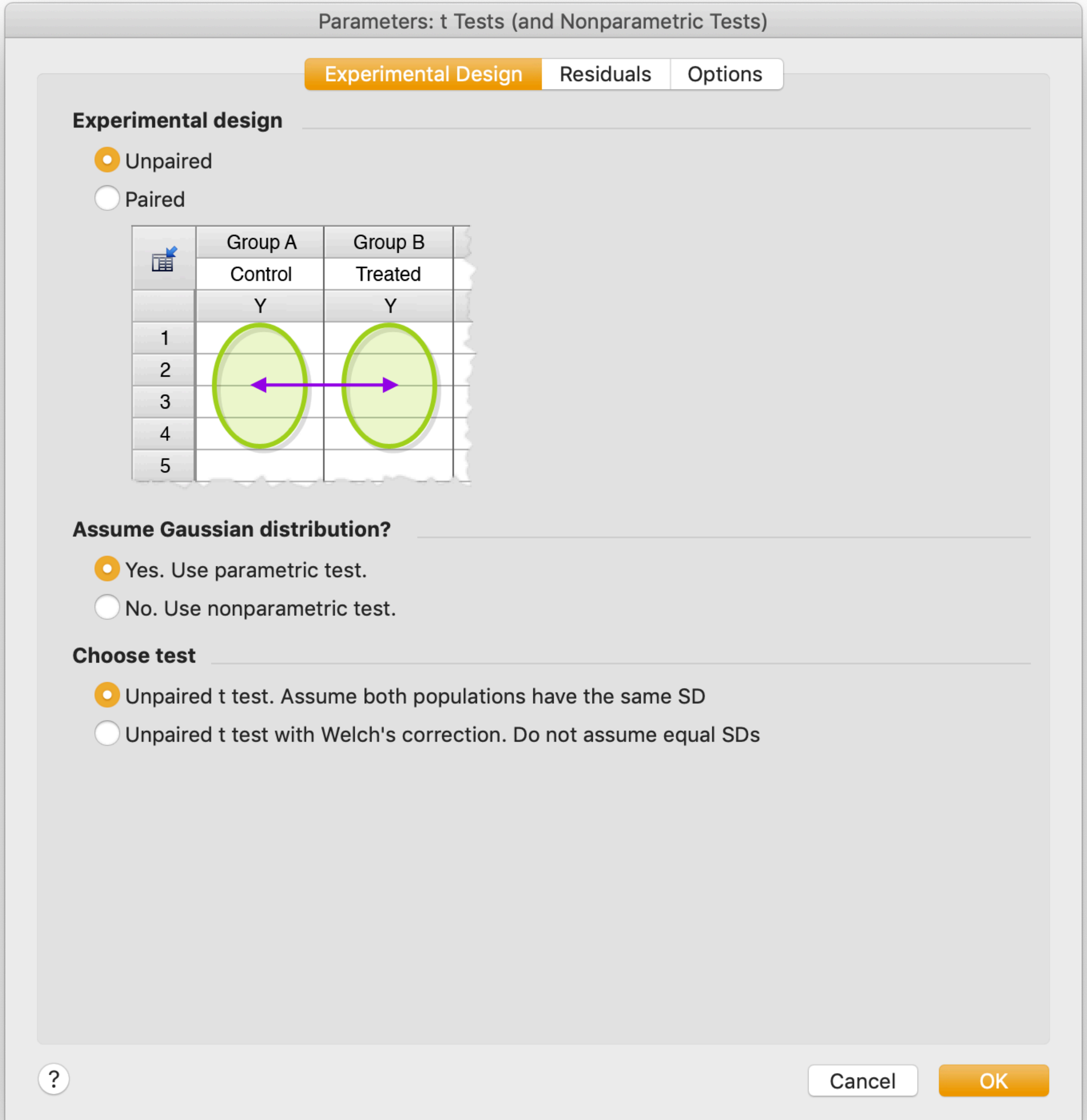
Test statistic:

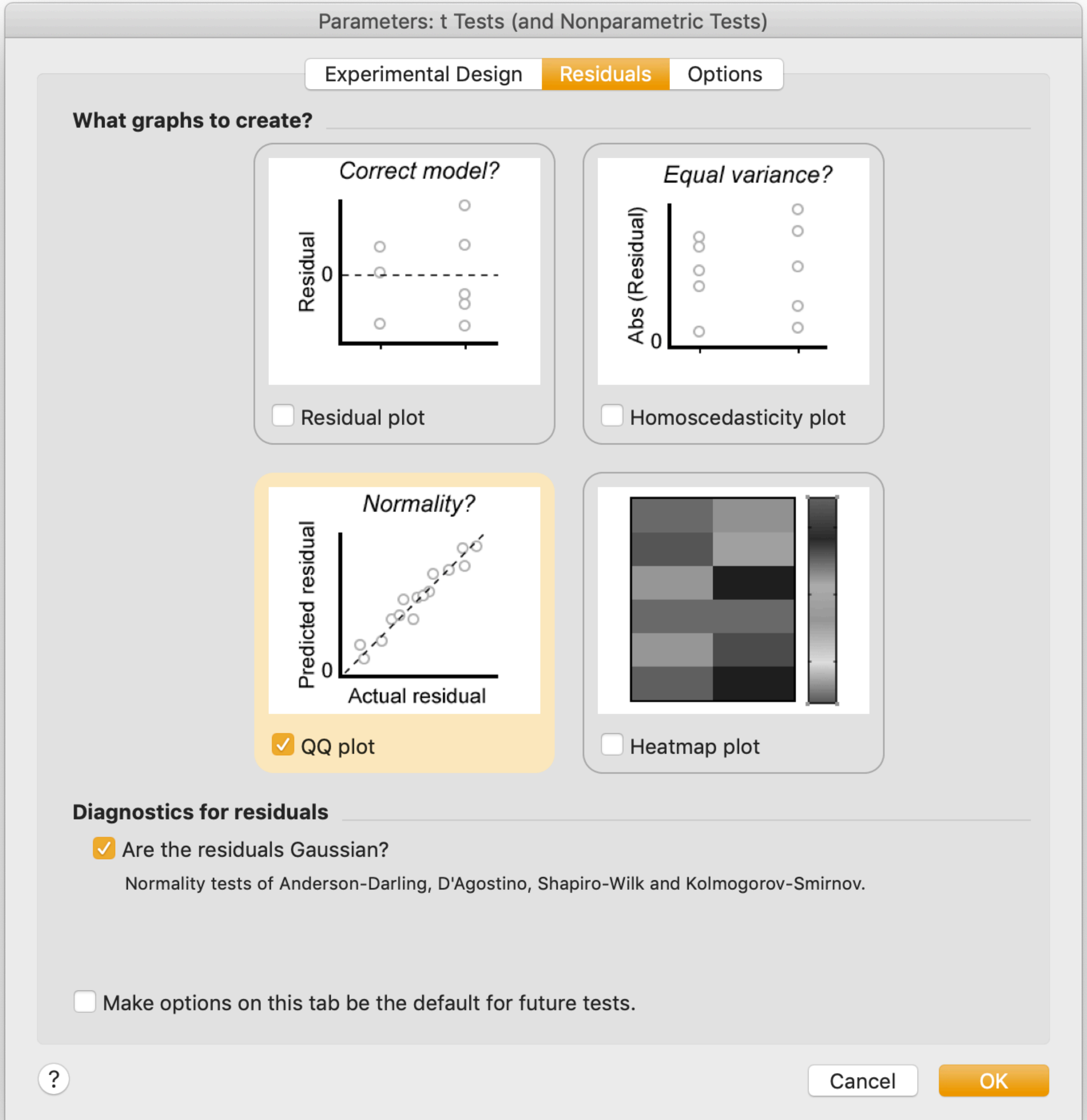
U (based on rank-order of xs and ys)

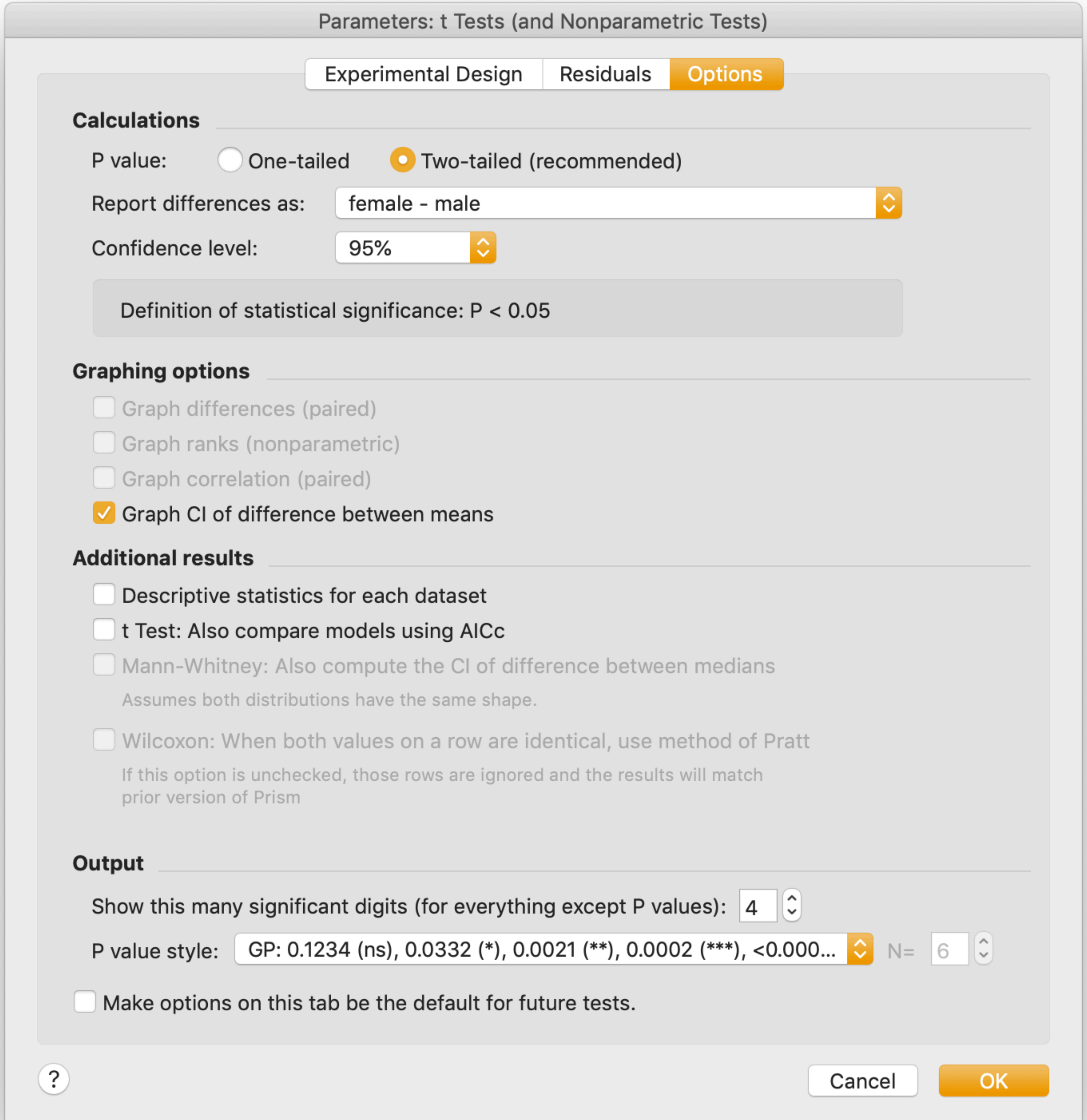


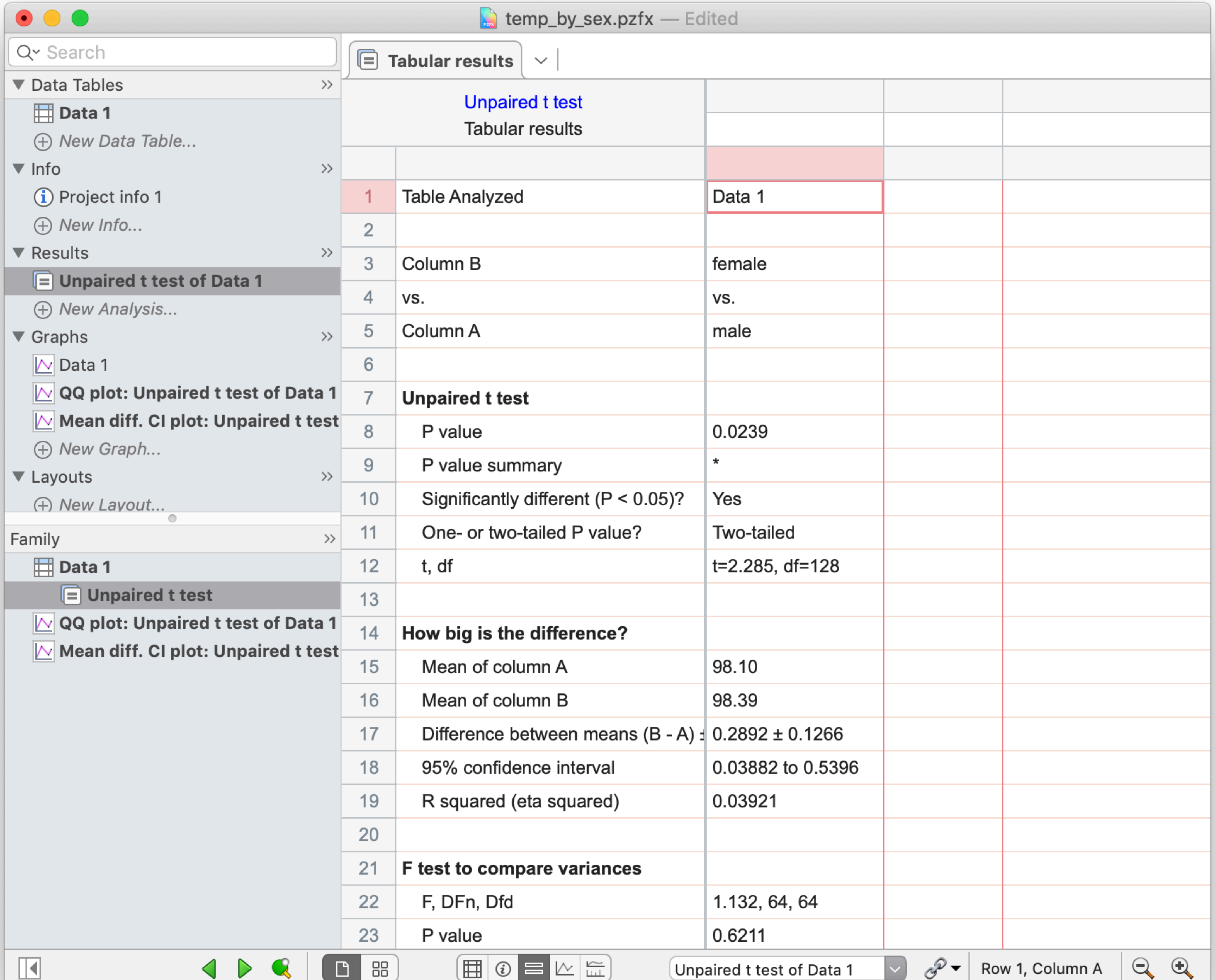
	Group A	Group B	Group C	Group D
	male	female	Title	Title
	Y	Y	Y	Y
1	96.3	96.4		
2	96.7	96.7		
3	96.9	96.8		
4	97.0	97.2		
5	97.1	97.2		
6	97.1	97.4		
7	97.1	97.6		
8	97.2	97.7		
9	97.3	97.7		
10	97.4	97.8		
11	97.4	97.8		
12	97.4	97.8		
13	97.4	97.9		
14	97.5	97.9		

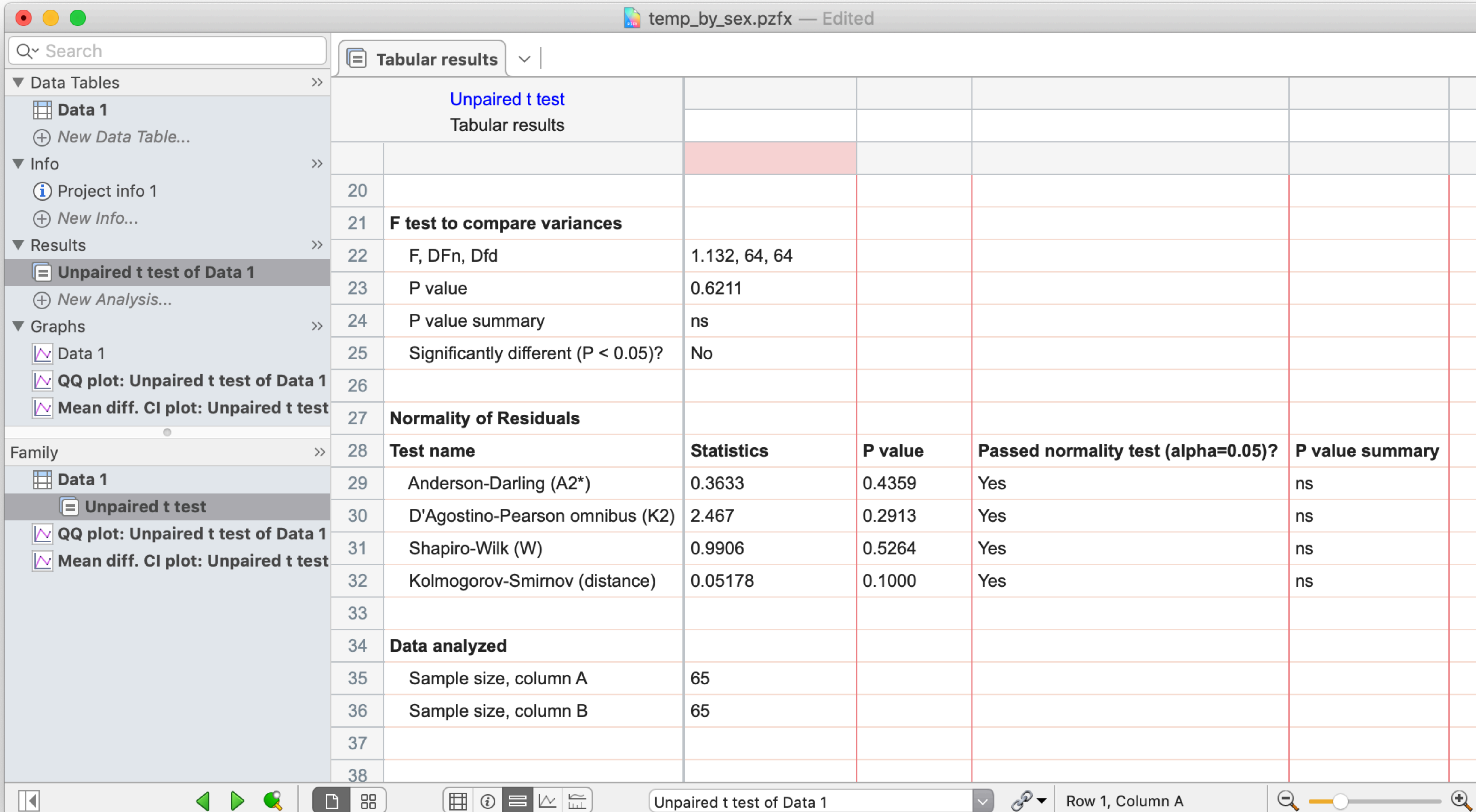

Create New Analysis
Data to analyze
Table: Data 1
Type of analysis
Which analysis?
Transform, Normalize...
XY analyses
Column analyses
t tests (and nonparametric tests)
One-way ANOVA (and nonparametric o...
One sample t and Wilcoxon test
Descriptive statistics
Normality and Lognormality Tests
Frequency distribution
ROC Curve
Bland-Altman method comparison
Identify outliers
Analyze a stack of P values
Grouped analyses
Contingency table analyses
Survival analyses
Parts of whole analyses
Multiple variable analyses
Nested analyses
Generate curve
Simulate data
Recently used
Analyze which data sets?
A:male
B:female
Select All
Deselect All
?
Cancel
OK


Parameters: t Tests (and Nonparametric Tests)
Experimental Design
Residuals
Options
Experimental design
Unpaired
Paired
Group A
Group B
Control
Treated
Y
Y
1
2
3
4
5
Assume Gaussian distribution?
Yes. Use parametric test.
No. Use nonparametric test.
Choose test
Unpaired t test. Assume both populations have the same SD
Unpaired t test with Welch's correction. Do not assume equal SDs
?
Cancel
OK

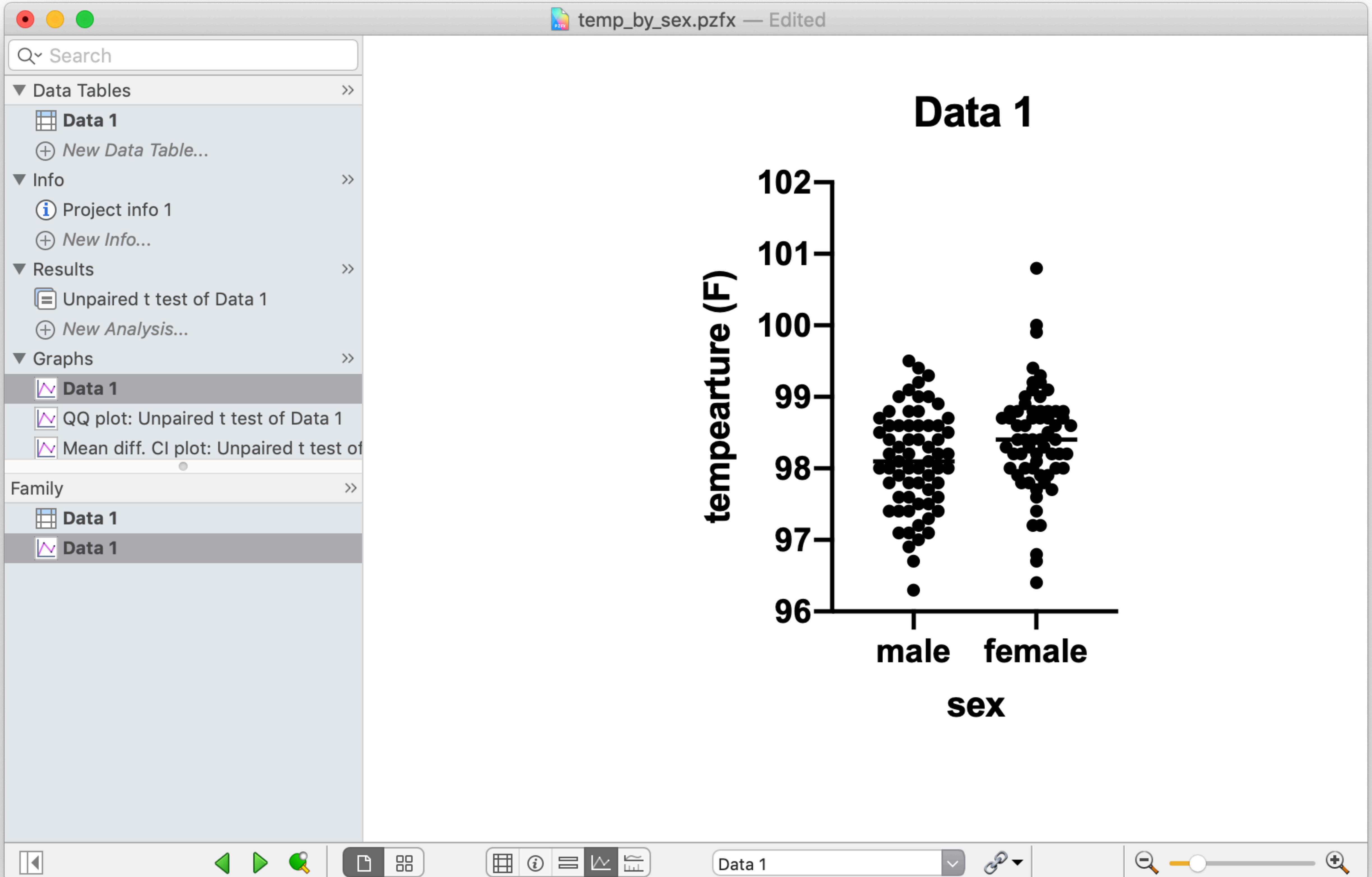

Parameters: t Tests (and Nonparametric Tests)
Experimental Design
Residuals
Options
What graphs to create?
Correct model?
Residual
0
Residual plot
Equal variance?
Abs (Residual)
0
Homoscedasticity plot
Normality?
Predicted residual
0
Actual residual
QQ plot
Heatmap plot
Diagnostics for residuals
Are the residuals Gaussian?
Normality tests of Anderson-Darling, D'Agostino, Shapiro-Wilk and Kolmogorov-Smirnov.
Make options on this tab be the default for future tests.
?
Cancel
OK

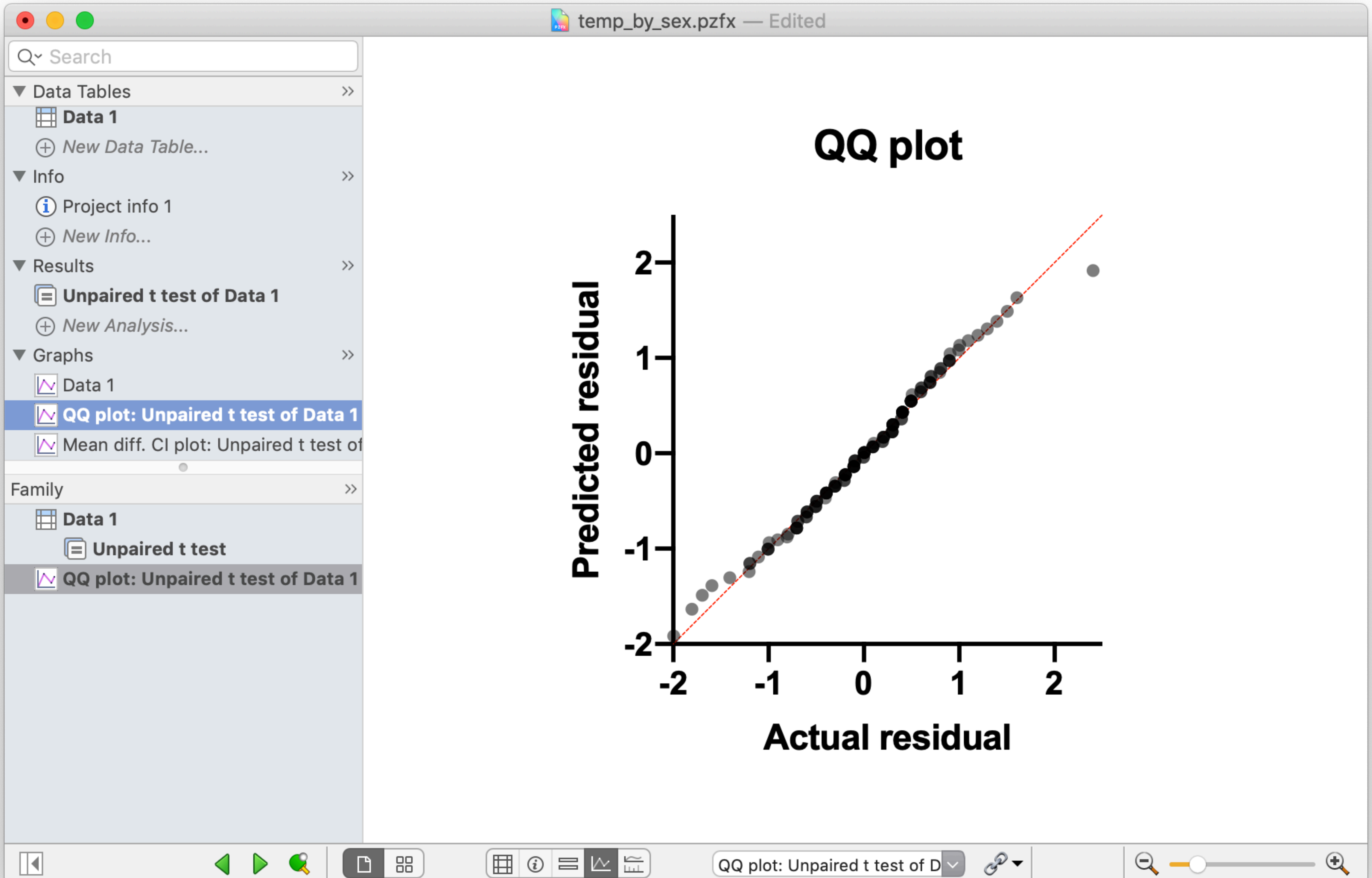

Parameters: t Tests (and Nonparametric Tests)
Experimental Design
Residuals
Options
Calculations
P value:
One-tailed
Two-tailed (recommended)
Report differences as:
female - male
Confidence level:
95%
Definition of statistical significance: P < 0.05
Graphing options
Graph differences (paired)
Graph ranks (nonparametric)
Graph correlation (paired)
Graph CI of difference between means
Additional results
Descriptive statistics for each dataset
t Test: Also compare models using AICc
Mann-Whitney: Also compute the CI of difference between medians
Assumes both distributions have the same shape.
Wilcoxon: When both values on a row are identical, use method of Pratt
If this option is unchecked, those rows are ignored and the results will match prior version of Prism
Output
Show this many significant digits (for everything except P values):
4
P value style:
GP: 0.1234 (ns), 0.0332 (*), 0.0021 (**), 0.0002 (***), <0.000...
N=
6
Make options on this tab be the default for future tests.
?
Cancel
OK

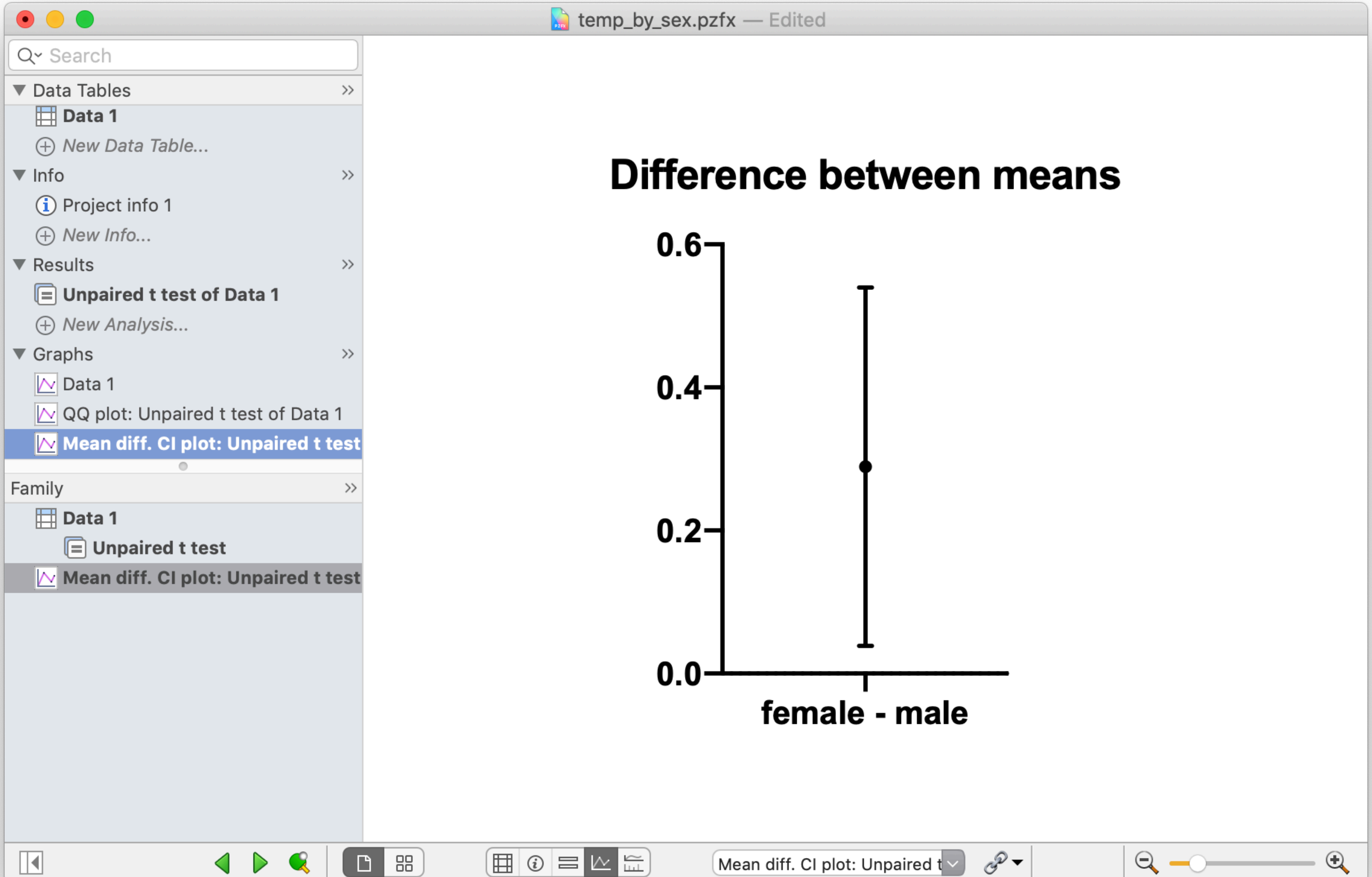

temp_by_sex.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Unpaired t test of Data 1
New Analysis...
Graphs
Data 1
QQ plot: Unpaired t test of Data 1
Mean diff. CI plot: Unpaired t test
New Graph...
Layouts
New Layout...
Family
Data 1
Unpaired t test
QQ plot: Unpaired t test of Data 1
Mean diff. CI plot: Unpaired t test
Tabular results
Unpaired t test
Tabular results
1 Table Analyzed | Data 1
2
3 Column B | female
4 vs. | vs.
5 Column A | male
6
7 Unpaired t test
8 P value | 0.0239
9 P value summary | *
10 Significantly different (P < 0.05)? | Yes
11 One- or two-tailed P value? | Two-tailed
12 t, df | t=2.285, df=128
13
14 How big is the difference?
15 Mean of column A | 98.10
16 Mean of column B | 98.39
17 Difference between means (B - A) ± | 0.2892 ± 0.1266
18 95% confidence interval | 0.03882 to 0.5396
19 R squared (eta squared) | 0.03921
20
21 F test to compare variances
22 F, DFn, Dfd | 1.132, 64, 64
23 P value | 0.6211
Unpaired t test of Data 1
Row 1, Column A



20					
21	**F test to compare variances**				
22	F, DFn, Dfd	1.132, 64, 64			
23	P value	0.6211			
24	P value summary	ns			
25	Significantly different (P < 0.05)?	No			
26					
27	**Normality of Residuals**				
28	**Test name**	**Statistics**	**P value**	**Passed normality test (alpha=0.05)?**	**P value summary**
29	Anderson-Darling (A2*)	0.3633	0.4359	Yes	ns
30	D'Agostino-Pearson omnibus (K2)	2.467	0.2913	Yes	ns
31	Shapiro-Wilk (W)	0.9906	0.5264	Yes	ns
32	Kolmogorov-Smirnov (distance)	0.05178	0.1000	Yes	ns
33					
34	**Data analyzed**				
35	Sample size, column A	65			
36	Sample size, column B	65			
37					
38					


temp_by_sex.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Unpaired t test of Data 1
New Analysis...
Graphs
Data 1
QQ plot: Unpaired t test of Data 1
Mean diff. CI plot: Unpaired t test of
Family
Data 1
Data 1
Data 1
102
101
100
99
98
97
96
tempearture (F)
male
female
sex
Data 1


temp_by_sex.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Unpaired t test of Data 1
New Analysis...
Graphs
Data 1
QQ plot: Unpaired t test of Data 1
Mean diff. CI plot: Unpaired t test of
Family
Data 1
Unpaired t test
QQ plot: Unpaired t test of Data 1
QQ plot
Predicted residual
Actual residual
-2
-1
0
1
2
QQ plot: Unpaired t test of D


temp_by_sex.pzfx — Edited
Search
Data Tables
Data 1
New Data Table...
Info
Project info 1
New Info...
Results
Unpaired t test of Data 1
New Analysis...
Graphs
Data 1
QQ plot: Unpaired t test of Data 1
Mean diff. CI plot: Unpaired t test
Family
Data 1
Unpaired t test
Mean diff. CI plot: Unpaired t test
Difference between means
0.6
0.4
0.2
0.0
female - male
Mean diff. CI plot: Unpaired t