What's Changed in How My Students Find (and Think About) P-values

Robin Lock
St. Lawrence University

USCOTS 2013 - Opening Session

A Historical Comparison

Traditional: p-value in the year 10 BC*

vs.

Modern: p-value today

*BC = Before Cobb

Example: Beer and Mosquitoes

Does consuming beer attract mosquitoes?

Experiment*:

- Volunteer were randomly assigned to drink a liter of either beer or water
- Mosquitoes were caught in traps as they approached the volunteers.

	n	mean	std. dev.
Beer	25	23.60	4.1
Water	18	19.22	3.7

H_0: $\mu_B = \mu_W$
H_0: $\mu_B > \mu_W$

*Lefvre, T., et. al., "Beer Consumption Increases Human Attractiveness to Malaria Mosquitoes," *PLoS ONE,* 2010; 5(3): e9546.

Traditional p-value

1. Pick a formula

$$t = \frac{\bar{x}_B - \bar{x}_W}{\sqrt{\frac{s_B^2}{n_B} + \frac{s_W^2}{n_W}}}$$

2. Plug & Chug

$$t = \frac{23.6 - 19.22}{\sqrt{\frac{4.1^2}{25} + \frac{3.7^2}{18}}} = 3.68$$

3. Pick a reference distribution

Where is my t-table?

0.0005 < p-value < 0.001

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	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	.692	.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	.691	.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	.690	.865	1.071	1.337	1.746	2.120	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
21	.686	.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	.686	.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	.685	.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768
24	.685	.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	.684	.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	.684	.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	.684	.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	.683	.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	.683	.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	.683	.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	.681	.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	.679	.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	.679	.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	.678	.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	.677	.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
1000	.675	.842	1.037	1.282	1.646	1.962	2.056	2.330	2.581	2.813	3.098	3.300
∞	.674	.841	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.091	3.291

d.f.

t=3.68

Oops! I forgot to check conditions...

Sample sizes are both less than 30, is the t-distribution even appropriate?

What if the data are heavily skewed?

...and, by the way, how does this method connect to the definition of a p-value?

p-value

The **p-value** is the proportion of samples, when H_0 is true, that would give results as (or more) extreme as the original sample.

Simulation Approach

"when H_0 is true" $\Rightarrow$ It doesn't matter whether the subject drank beer or water

Create randomization samples (under H_0) by randomly re-assigning the beer/water labels to the 43 mosquito counts.

Find the difference, $\bar{x}_B - \bar{x}_W$, for each sample.

Physical Simulation

- Put the 43 mosquito counts on cards.
- Shuffle and deal cards into two piles (25 beer and 18 water).
- Compute the difference in means.
- Repeat MANY times.

We Need Some Technology!

StatKey

www.lock5stat.com/statkey

StatKey to accompany *Statistics: Unlocking the Power of Data*

by Lock, Lock, Lock, Lock, and Lock

Descriptive Statistics and Graphs	Bootstrap Confidence Intervals	Randomization Hypothesis Tests
One Quantitative Variable	CI for Single Mean, Median, St.Dev.	Test for Single Mean
One Categorical Variable	CI for Single Proportion	Test for Single Proportion
One Quantitative and One Categorical Variable	CI for Difference In Means	Test for Difference in Means
Two Categorical Variables	CI for Difference In Proportions	Test for Difference In Proportions
Two Quantitative Variables	CI for Slope, Correlation	Test for Slope, Correlation

Sampling Distributions	Mean	Proportion

Theoretical Distributions	Normal	t	χ^2	F

More Advanced Randomization Tests	χ^2 Goodness-of-Fit	χ^2 Test for Association	ANOVA for Difference in Means	ANOVA for Regression

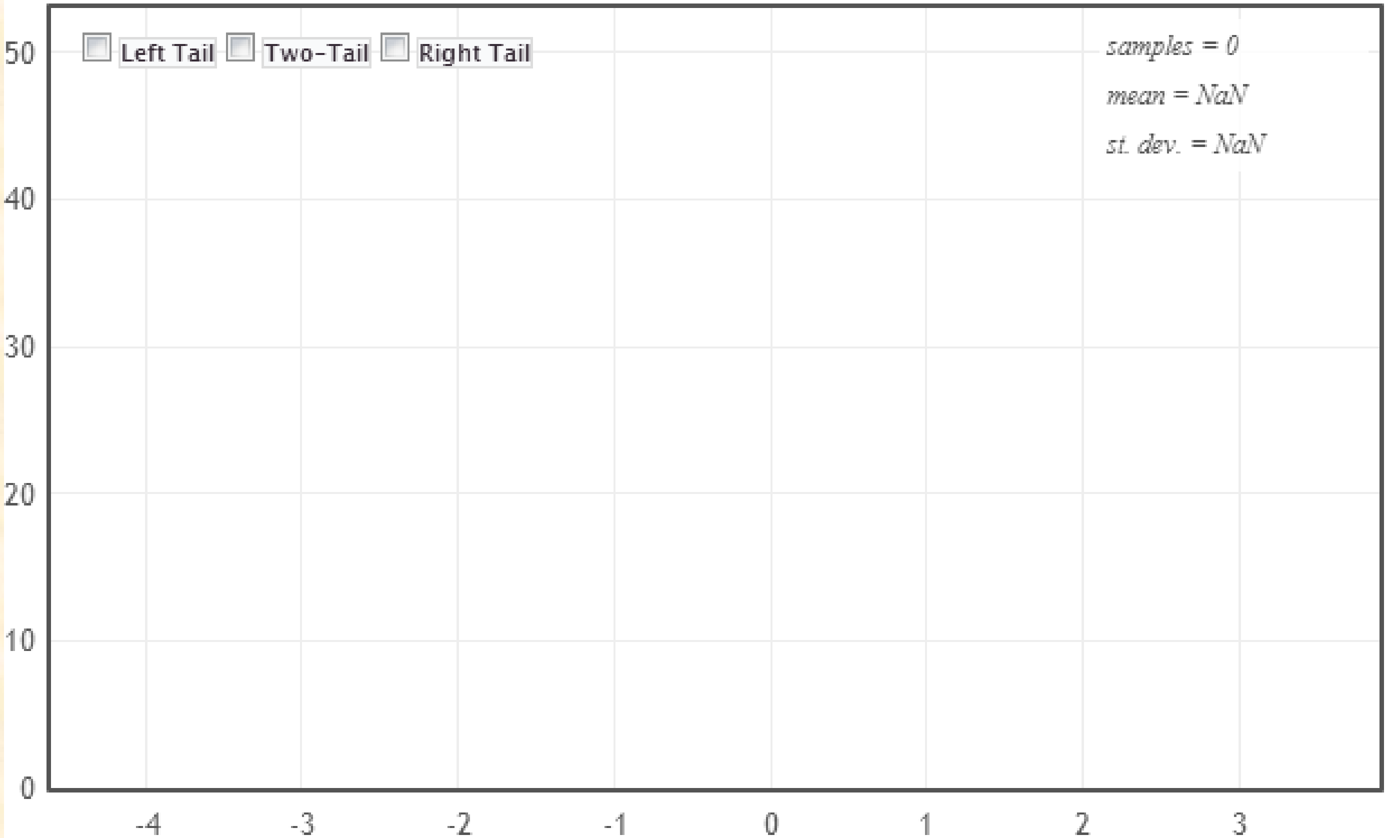
StatKey Randomization Test for a Difference in Means

Mosquitoes (Beer vs Water) | Show Data Table | Edit Data

Randomization method: Reallocate Groups

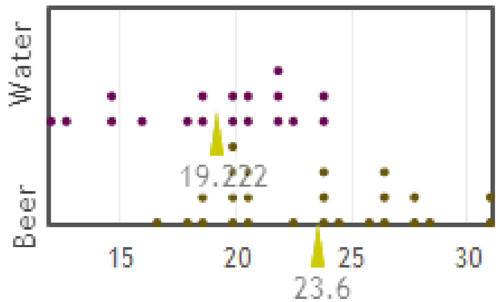
Generate 1 Sample | Generate 10 Samples | Generate 100 Samples | Generate 1000 Samples | Reset Plot

Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$



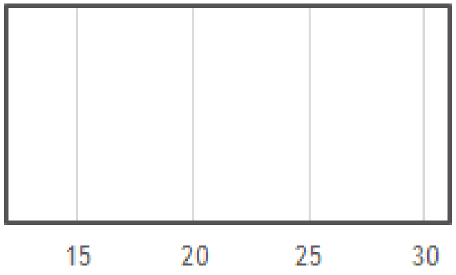
Original Sample

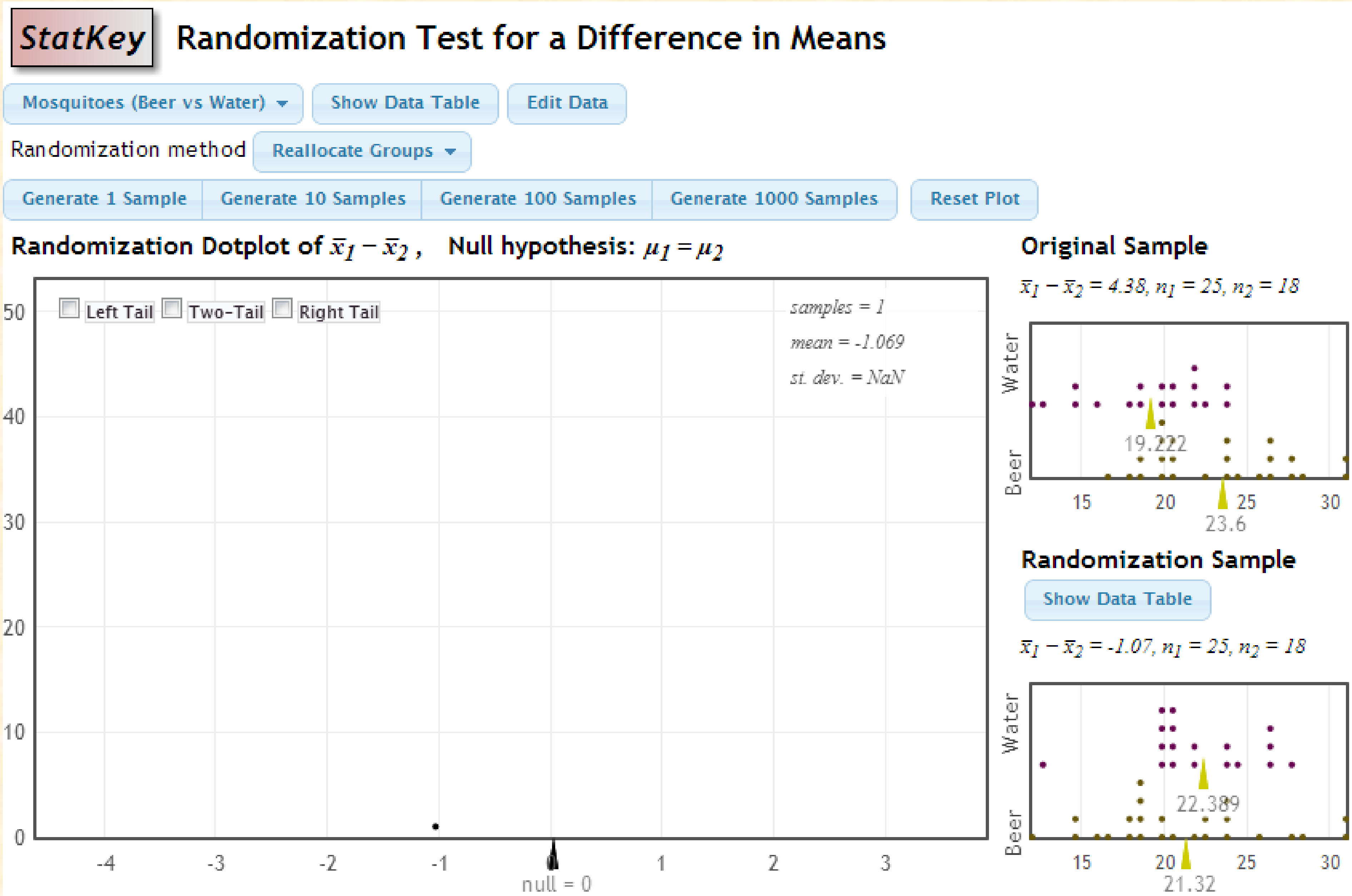
$\bar{x}_1 - \bar{x}_2 = 4.38$, $n_1 = 25$, $n_2 = 18$

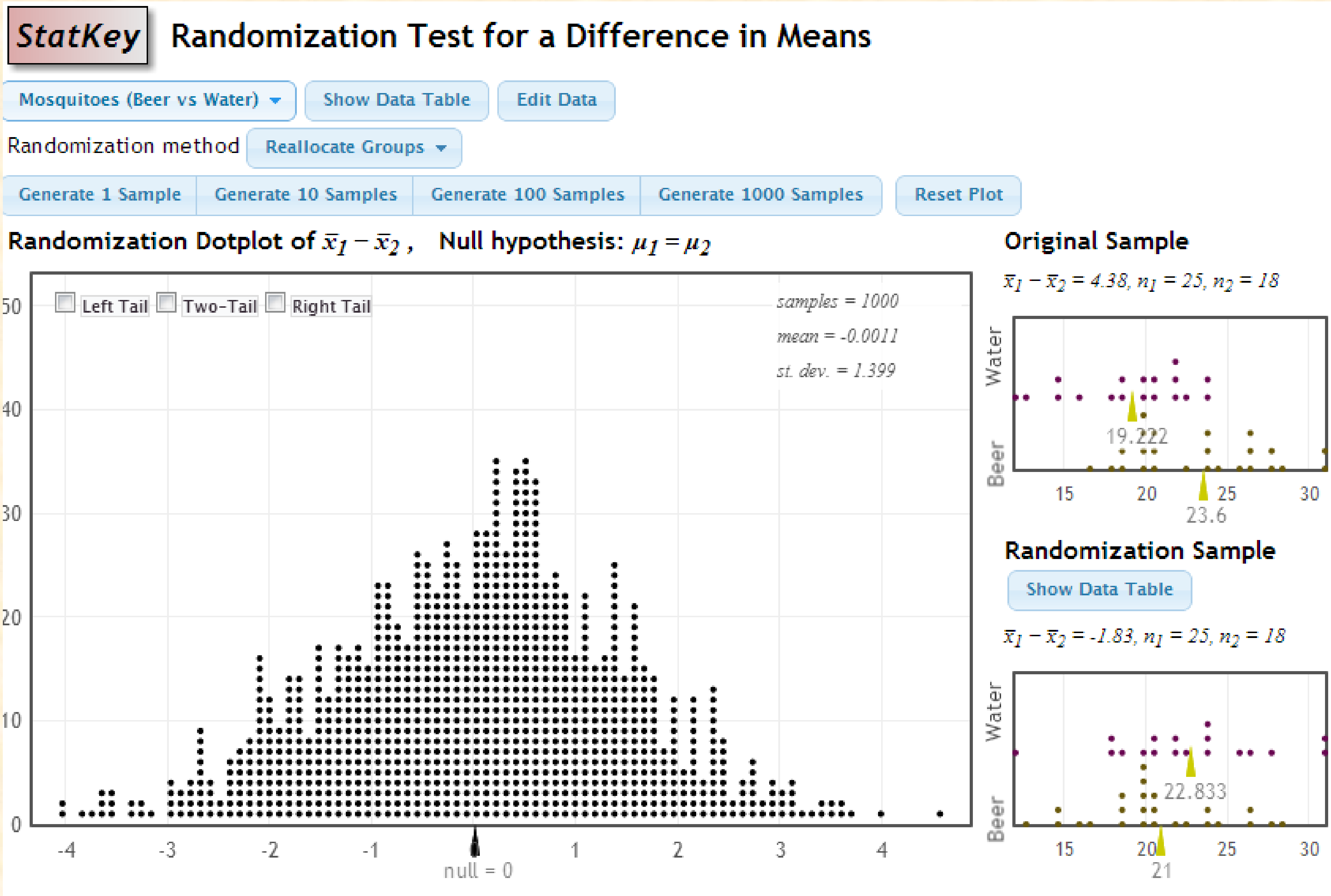


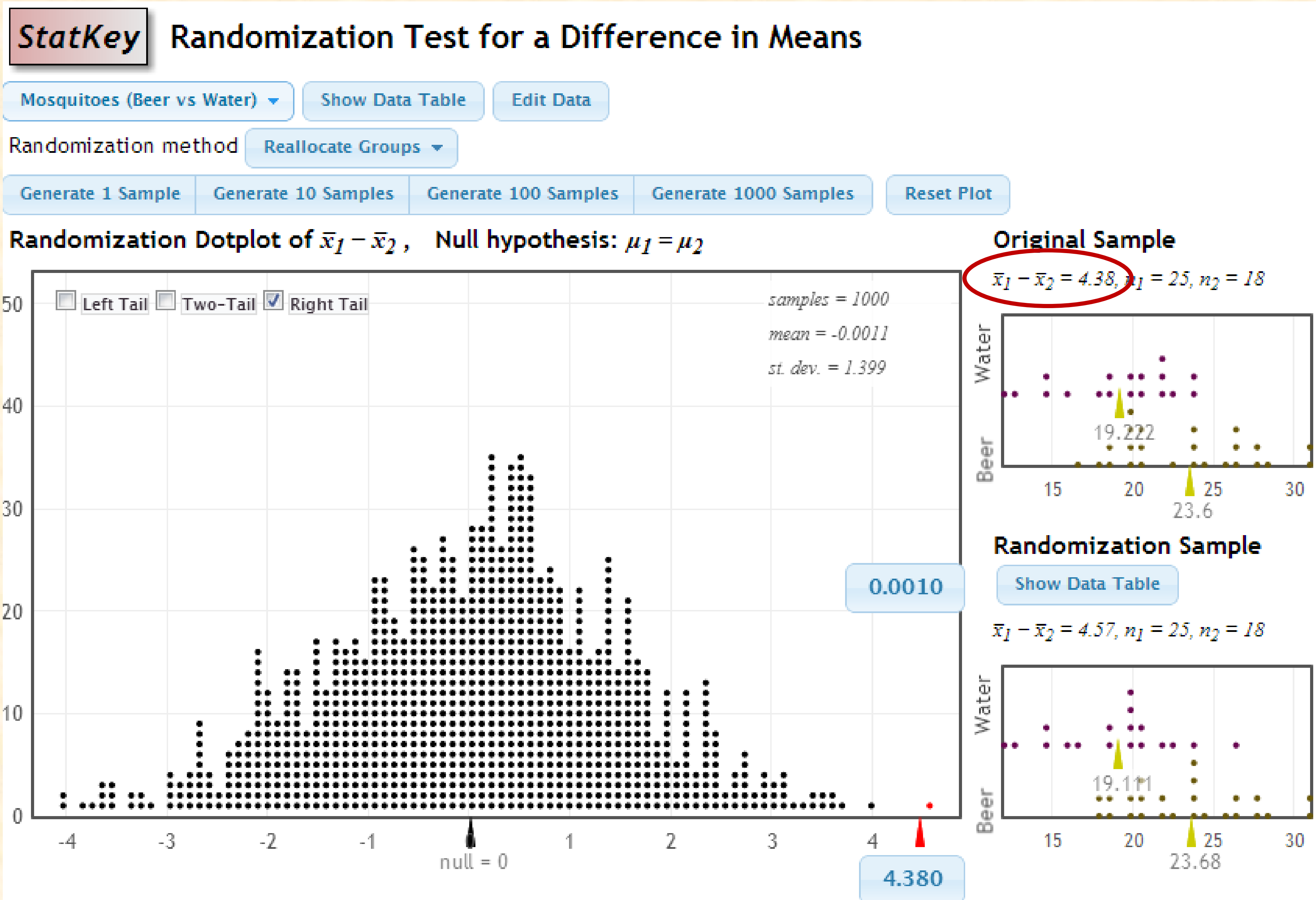
Randomization Sample

Show Data Table




StatKey
Randomization Test for a Difference in Means
Mosquitoes (Beer vs Water)
Show Data Table
Edit Data
Randomization method
Reallocate Groups
Generate 1 Sample
Generate 10 Samples
Generate 100 Samples
Generate 1000 Samples
Reset Plot
Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$
Left Tail
Two-Tail
Right Tail
samples = 1
mean = -1.069
st. dev. = NaN
50
40
30
20
10
0
-4
-3
-2
-1
0
1
2
3
null = 0
Original Sample
$\bar{x}_1 - \bar{x}_2 = 4.38, n_1 = 25, n_2 = 18$
Water
Beer
19.222
23.6
15
20
25
30
Randomization Sample
Show Data Table
$\bar{x}_1 - \bar{x}_2 = -1.07, n_1 = 25, n_2 = 18$
Water
Beer
22.389
21.32
15
20
25
30


StatKey
Randomization Test for a Difference in Means
Mosquitoes (Beer vs Water)
Show Data Table
Edit Data
Randomization method
Reallocate Groups
Generate 1 Sample
Generate 10 Samples
Generate 100 Samples
Generate 1000 Samples
Reset Plot
Randomization Dotplot of $\bar{x}_1 - \bar{x}_2$, Null hypothesis: $\mu_1 = \mu_2$
Left Tail
Two-Tail
Right Tail
samples = 1000
mean = -0.0011
st. dev. = 1.399
null = 0
Original Sample
$\bar{x}_1 - \bar{x}_2 = 4.38, n_1 = 25, n_2 = 18$
Water
Beer
19.222
23.6
Randomization Sample
Show Data Table
$\bar{x}_1 - \bar{x}_2 = -1.83, n_1 = 25, n_2 = 18$
Water
Beer
22.833
21



"as extreme as the original sample"

Background Required

- Random shuffle
- Compute sample means
- Dotplot
- Find a proportion by counting

Questions?

- What about other parameters or different hypotheses?
- What about checking conditions?
- What about confidence intervals?
- What was that address for StatKey?

www.lock5stat.com/statkey

StatKey

to accompany *Statistics: Unlocking the Power of Data*

by Lock, Lock, Lock, Lock, and Lock

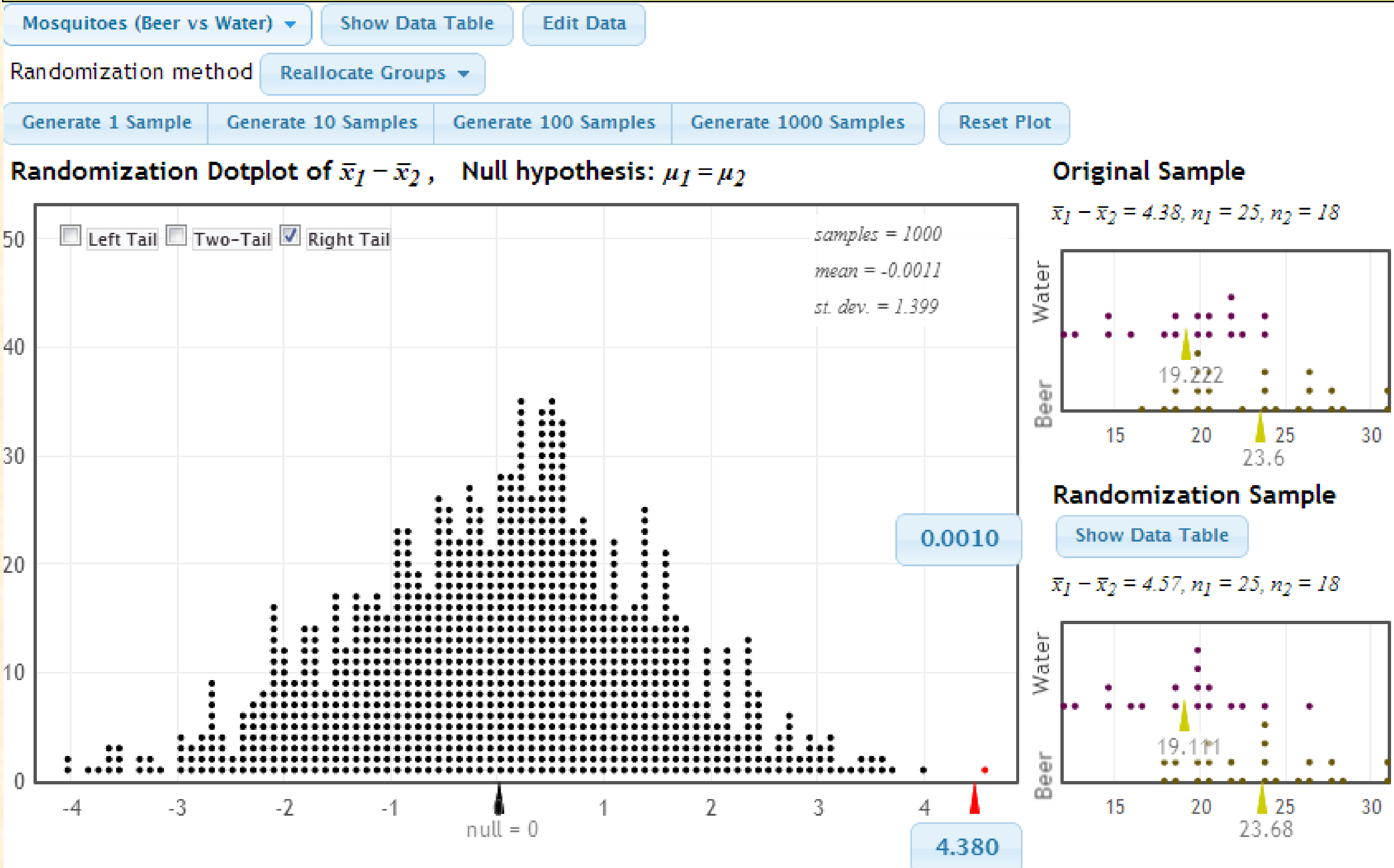
Descriptive Statistics and Graphs	Bootstrap Confidence Intervals	Randomization Hypothesis Tests
One Quantitative Variable	CI for Single Mean, Median, St.Dev.	Test for Single Mean
One Categorical Variable	CI for Single Proportion	Test for Single Proportion
One Quantitative and One Categorical Variable	CI for Difference In Means	Test for Difference in Means
Two Categorical Variables	CI for Difference In Proportions	Test for Difference In Proportions
Two Quantitative Variables	CI for Slope, Correlation	Test for Slope, Correlation

Sampling Distributions	Mean	Proportion

Theoretical Distributions	Normal	t	χ^2	F

More Advanced Randomization Tests	χ^2 Goodness-of-Fit	χ^2 Test for Association	ANOVA for Difference in Means	ANOVA for Regression

p-value = proportion of samples, when H_0 is true, that are as (or more) extreme as the original sample.



p-value = proportion of samples, when H_0 is true, that are as (or more) extreme as the original sample.