Math 3081 (Probability and Statistics)

Lecture #25 of 27 $\sim$ August 17th, 2021

Matched Pairs and Robustness of t Tests

- Matched Pairs
- Robustness of t Tests
- More Examples of t Tests

This material represents §5.1.4-5.1.5 from the course notes, and problems 10-14 from WeBWorK 7.

Recall, I

Last time, we discussed Student's equal-variances two-sample t-test for comparing the means of two independent, normally-distributed populations A and B with unknown standard deviations that are assumed to be equal.

- For this test, our null hypothesis is of the form H_0: $\mu_A - \mu_B = c$ for some constant c that is our hypothesized value for the difference of the means (usually 0).
- We take the test statistic $t = \dfrac{(\hat{\mu}_A - \hat{\mu}_B) - c}{S_{\text{pool}}\sqrt{\dfrac{1}{n_A} + \dfrac{1}{n_B}}}$, where $S_{\text{pool}} = \sqrt{\dfrac{(n_A - 1)S_A^2 + (n_B - 1)S_B^2}{n_A + n_B - 2}}$ is the pooled standard deviation estimate.
- The test statistic is t-distributed with $n_A + n_B - 2$ degrees of freedom.

Recall, II

We also discussed Welch's unequal-variances two-sample t-test for comparing the means of two independent, normally-distributed populations A and B with unknown standard deviations, where we do not assume the population standard deviations are equal.

- Our null hypothesis is again of the form H_0: $\mu_A - \mu_B = c$ for some constant c that is our hypothesized value for the difference of the means (usually 0).
- The test statistic is $t = \dfrac{(\hat{\mu}_A - \hat{\mu}_B) - c}{S_{\text{unpool}}}$, where we use the unpooled standard deviation $S_{\text{unpool}} = \sqrt{\dfrac{S_A^2}{n_A} + \dfrac{S_B^2}{n_B}}$.
- The test statistic is (approximately) t-distributed with $\dfrac{(S_A^2/n_A + S_B^2/n_B)^2}{\frac{1}{n_A - 1}(S_A^2/n_A)^2 + \frac{1}{n_B - 1}(S_B^2/n_B)^2}$ degrees of freedom.

~~Two~~ One-Sample t Tests: Matched Pairs, I

We now discuss one additional scenario involving t tests and the comparison of two samples, involving matched pairs.

- In matched-pairs comparisons, we are comparing the means of two sets of paired data.
- A common situation is to make a before-and-after comparison of measurements taken before applying a treatment to measurements taken after applying the treatment; we want to know if the treatment affected the average outcome.
- Although this scenario involves two data sets, the matched-pairs design means that the initial and later measurements will be correlated, so it is not appropriate to use a two-sample t test.
- Instead, what we do is compute the difference in the results (for each individual), and use a one-sample t test to compare the average outcome to 0.

Matched Pairs, II

Example: An instructor has 6 students take a pre-assessment, complete a study module, and then a post-assessment. The results are below. Test at the 9%, 1%, and 0.3% significance levels if the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

Matched Pairs, II

Example: An instructor has 6 students take a pre-assessment, complete a study module, and then a post-assessment. The results are below. Test at the 9%, 1%, and 0.3% significance levels if the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

- Here, we have matched-pair data, because the measurements of the scores are coming from the same students. Since the values in the samples are not independent, but come from matched pairs, we want to use a one-sample t test here.

Matched Pairs, III

Example: Test at the 9%, 1%, and 0.3% significance levels if the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

Matched Pairs, III

Example: Test at the 9%, 1%, and 0.3% significance levels if the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

- Our hypotheses are $H_0 : \mu_{\text{post}} = \mu_{\text{pre}}$ and $H_a : \mu_{\text{post}} > \mu_{\text{pre}}$, which we can rephrase in terms of the difference in means $\mu_{\text{diff}} = \mu_{\text{post}} - \mu_{\text{pre}}$ as H_0: $\mu_{\text{diff}} = 0$ and $H_a : \mu_{\text{diff}} > 0$.
- Our test statistic is the difference in means μ_{diff}, which will be t-distributed with $6 - 1 = 5$ degrees of freedom.

Matched Pairs, IV

Example: Test at the 9%, 1%, and 0.3% significance levels if the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

Matched Pairs, IV

Example: Test at the 9%, 1%, and 0.3% significance levels if the students' scores improved after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

- Our sample data set consists of the six differences of scores $\{13, 17, 7, -1, 30, 12\}$, with mean $\hat{\mu}_{\text{diff}} = 13$ and sample standard deviation $S = 10.3730$, and the value of the sample statistic is $t = \dfrac{\hat{\mu}_{\text{diff}} - 0}{S/\sqrt{n}} = 3.0698$.
- Thus, the p-values is $P(T_5 \geq 3.0698) = 0.01389$.
- Since the p-value is below 9% we reject the null hypothesis at that significance level, but since it is above 1% and 0.3% we fail to reject at those significance levels.

Matched Pairs, V

Example: Find a 95% confidence interval for the average improvement in a student's test score after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

Matched Pairs, V

Example: Find a 95% confidence interval for the average improvement in a student's test score after studying:

Student	A	B	C	D	E	F
Pre-study	61	71	90	81	55	81
Post-study	74	88	97	80	85	93

- We can construct a confidence interval using the fact that the test statistic $\frac{\mu_{\text{diff}}}{S/\sqrt{n}}$ will be t-distributed with $6-1=5$ degrees of freedom.
- Since the sample mean is $\hat{\mu}_{\text{diff}} = 13$ with sample standard deviation $S = 10.3730$, and the t-statistic $t_{\alpha/2,df}$ for $\alpha = 95\%$ and $df = 5$ is $t = 2.5706$, the desired confidence interval is $13 \pm 2.5706 \cdot 10.3730/\sqrt{5} = (2.1142, 23.8858)$.

Matched Pairs, VI

Example: To determine whether a new drug lowers blood serum LDL cholesterol levels, 300 patients are given the drug, and their cholesterol levels are measured before starting a course of the drug, and then again after 5 years of taking the drug. The starting LDL levels averaged 173.86 mg/dL with a sample standard deviation of 29.75 mg/dL, while the ending LDL levels averaged 168.25 mg/dL with a sample standard deviation of 30.20 mg/dL. The sample standard deviation for the differences for each patient was 39.44 mg/dL. Assume all relevant quantities are normally distributed.

1. Test at the 1% significance level whether LDL levels were lower after 5 years.
2. Give 99% confidence intervals for the starting LDL level, the ending LDL level, and the difference in LDL levels.
3. Explain why it is not possible to conclude from these calculations that the drug lowered LDL levels.

Matched Pairs, VII

Example: 300 patients have starting LDL levels averaging 173.86 mg/dL with a sample standard deviation of 29.75 mg/dL, and the ending LDL levels averaging 168.25 mg/dL with a sample standard deviation of 30.20 mg/dL. The sample standard deviation for the differences for each patient was 39.44 mg/dL.

1. Test at the 1% significance level whether LDL levels were lower after 5 years.

Matched Pairs, VII

Example: 300 patients have starting LDL levels averaging 173.86 mg/dL with a sample standard deviation of 29.75 mg/dL, and the ending LDL levels averaging 168.25 mg/dL with a sample standard deviation of 30.20 mg/dL. The sample standard deviation for the differences for each patient was 39.44 mg/dL.

1. Test at the 1% significance level whether LDL levels were lower after 5 years.

- Here, we have matched-pair data, because the values are taken from the same patients before and after, so we want to use a one-sample t test on the differences [end] - [start].
- Our hypotheses are $H_0 : \mu_{\text{diff}} = 0$ and $H_a : \mu_{\text{diff}} > 0$.
- Our sample data set consists of the 600 differences of LDL levels (so $df = 299$), with mean $\hat{\mu}_{\text{diff}} = 173.86 - 168.25 = 5.61$ mg/dL and sample standard deviation $S_{\text{diff}} = 39.44$ mg/dL.

Matched Pairs, VIII

Example: 300 patients have starting LDL levels averaging 173.86 mg/dL with a sample standard deviation of 29.75 mg/dL, and the ending LDL levels averaging 168.25 mg/dL with a sample standard deviation of 30.20 mg/dL. The sample standard deviation for the differences for each patient was 39.44 mg/dL.

1. Test at the 1% significance level whether LDL levels were lower after 5 years.

- We have $df = 599$, $\hat{\mu}_{\text{diff}} = 175.28 - 168.25 = 5.61$ mg/dL, and $S_{\text{diff}} = 39.44$ mg/dL.
- The test statistic is $\dfrac{5.61 \text{ mg/dL}}{(39.44 \text{ mg/dL})/\sqrt{300}} = 2.4637$.
- Thus, the p-value is $P(T_{299} \geq 2.4637) = 0.00702$.
- Since the p-value is below 1%, we reject the null hypothesis and conclude that the LDL levels were indeed lowered.

Matched Pairs, IX

Example: 300 patients have starting LDL levels averaging 173.86 mg/dL with a sample standard deviation of 29.75 mg/dL, and the ending LDL levels averaging 168.25 mg/dL with a sample standard deviation of 30.20 mg/dL. The sample standard deviation for the differences for each patient was 39.44 mg/dL.

2. Give 99% confidence intervals for the starting LDL level, the ending LDL level, and the difference in LDL levels.

Matched Pairs, IX

Example: 300 patients have starting LDL levels averaging 173.86 mg/dL with a sample standard deviation of 29.75 mg/dL, and the ending LDL levels averaging 168.25 mg/dL with a sample standard deviation of 30.20 mg/dL. The sample standard deviation for the differences for each patient was 39.44 mg/dL.

2. Give 99% confidence intervals for the starting LDL level, the ending LDL level, and the difference in LDL levels.

- Here $df = 299$ so the t-statistic for a 99% CI is $t = 1.9679$.
- Start: $(173.86\ \text{mg/dL}) \pm 1.9639 \cdot (29.75\ \text{mg/dL})/\sqrt{300}$ $= (170.48\ \text{mg/dL}, 177.24\ \text{mg/dL})$.
- End: $(168.25\ \text{mg/dL}) \pm 1.9639 \cdot (30.20\ \text{mg/dL})/\sqrt{300}$ $= (164.82\ \text{mg/dL}, 171.68\ \text{mg/dL})$.
- Difference: $(-5.61\ \text{mg/dL}) \pm 1.9639 \cdot (39.44\ \text{mg/dL})/\sqrt{300}$ $= (-10.09\ \text{mg/dL}, -1.13\ \text{mg/dL})$.
- Note that although the 99% CIs for the starting and ending LDL levels, the 99% CI for the difference does not contain 0.

Matched Pairs, X

Example: LDL example.

3. Explain why it is not possible to conclude from these calculations that the drug lowered LDL levels.

Matched Pairs, X

Example: LDL example.

3. Explain why it is not possible to conclude from these calculations that the drug lowered LDL levels.

- The point here is not statistical in nature: although LDL levels were lowered, this study does not establish that the lowering was caused by the drug.
- To establish that, we would need to use a double-blind protocol that also includes a placebo group, since the results could be caused by the placebo effect.
- They could also an effect of participant selection: if people with high LDL levels are recruited by the study, one would expect to see regression to the mean.

Matched Pairs, XI

Regression to the mean is a very common phenomenon that can be observed in all sorts of places: sports statistics, medical trials, heights of parents and children, successful stocks, etc.

- To explain: there is variation in the measured quantity, so some high performers are actually closer to normal performers that were just measured on a "high" day. When measured again, they are much more likely to tend to have a lower measurement.
- Since we are only selecting high performers, our sample will be biased towards selecting elements that had lucky variation upwards, and in the future they will be more likely to land at a lower value.
- The same thing occurs in the reverse direction if we select low performers: their future average will tend to move upwards.

Robustness of t Tests, I

Over the last six lectures or so, we talked about z tests, the t distribution, and t tests.

- All of our discussion of z tests and t tests has been predicated on the assumption that the underlying populations we are studying are normally distributed.
- In reality, except for very rare examples arising in physics with phenomena having exact theoretical models, no population is precisely normally distributed.
- It is therefore important to study how well the tests we have developed will perform in situations where the underlying distributions are not exactly normal.
- What we are investigating is called robustness: the accuracy of the tests when applied to distributions that are not exactly the ones predicted by the model.

Robustness of t Tests, II

Our concern is similar to that which motivated our discussion of the t distribution and t tests.

- Specifically, when we do not know the standard deviation, we could simply have tried using z tests but with S in place of σ. The resulting test would then not be exact, but we could hope that it is fairly close.
- As we have explained, with small samples using a z place instead of a t test will generally be much less accurate (in the sense that the type I and type II error probabilities will generally be much larger).
- However, with large samples (e.g., n around 100 or more) then the difference between the standard normal distribution and the t distribution is negligible, and so using a z test in place of a t test in such situations does not introduce much error.

Robustness of t Tests, III

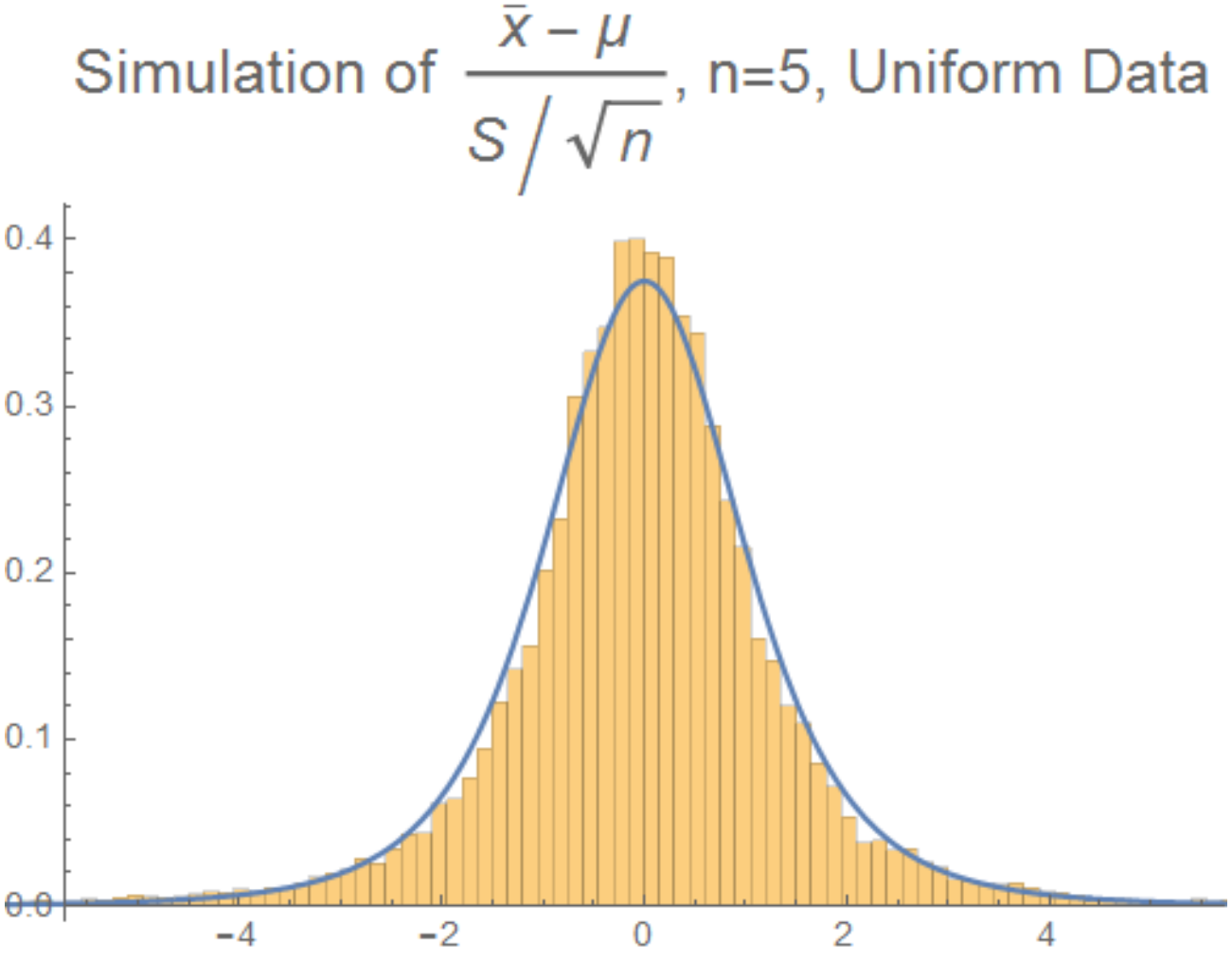
In principle, if we had a different underlying distribution (e.g., a uniform distribution), we could develop analogues of the z test and t test specifically for that underlying distribution.

- In fact there are many other statistical tests that have been developed precisely to allow accurate study of data sets that have very non-normally-shaped distributions.
- What we would like to know, though, is how necessary it is to expend this effort to develop a different test statistic for different types of data distributions.

It turns out that the t test is actually fairly robust, in that it performs fairly well even with distributions that are moderately non-normal.

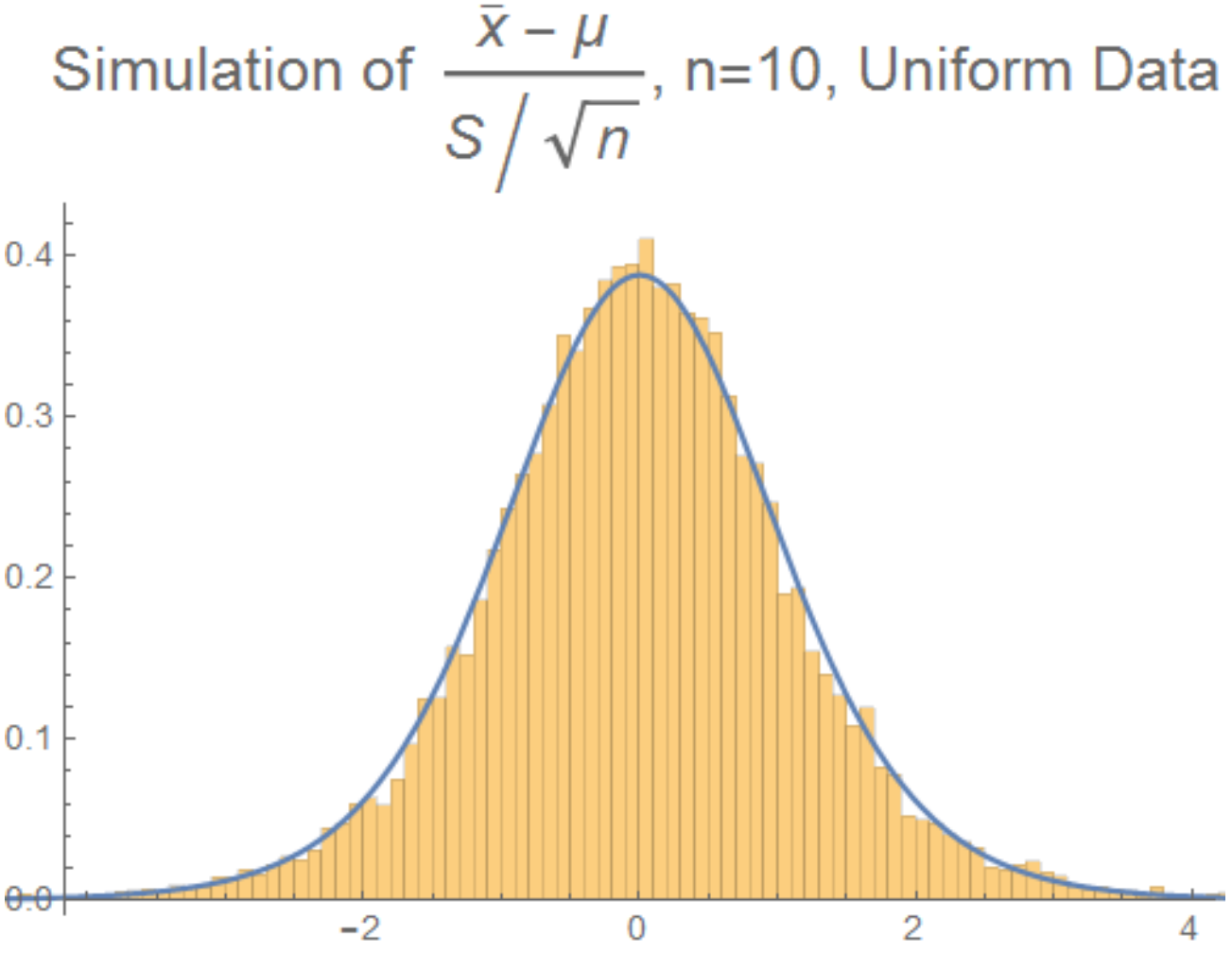
Robustness of t Tests, IV

Here are simulations of the t-statistic for sampling a uniform distribution:



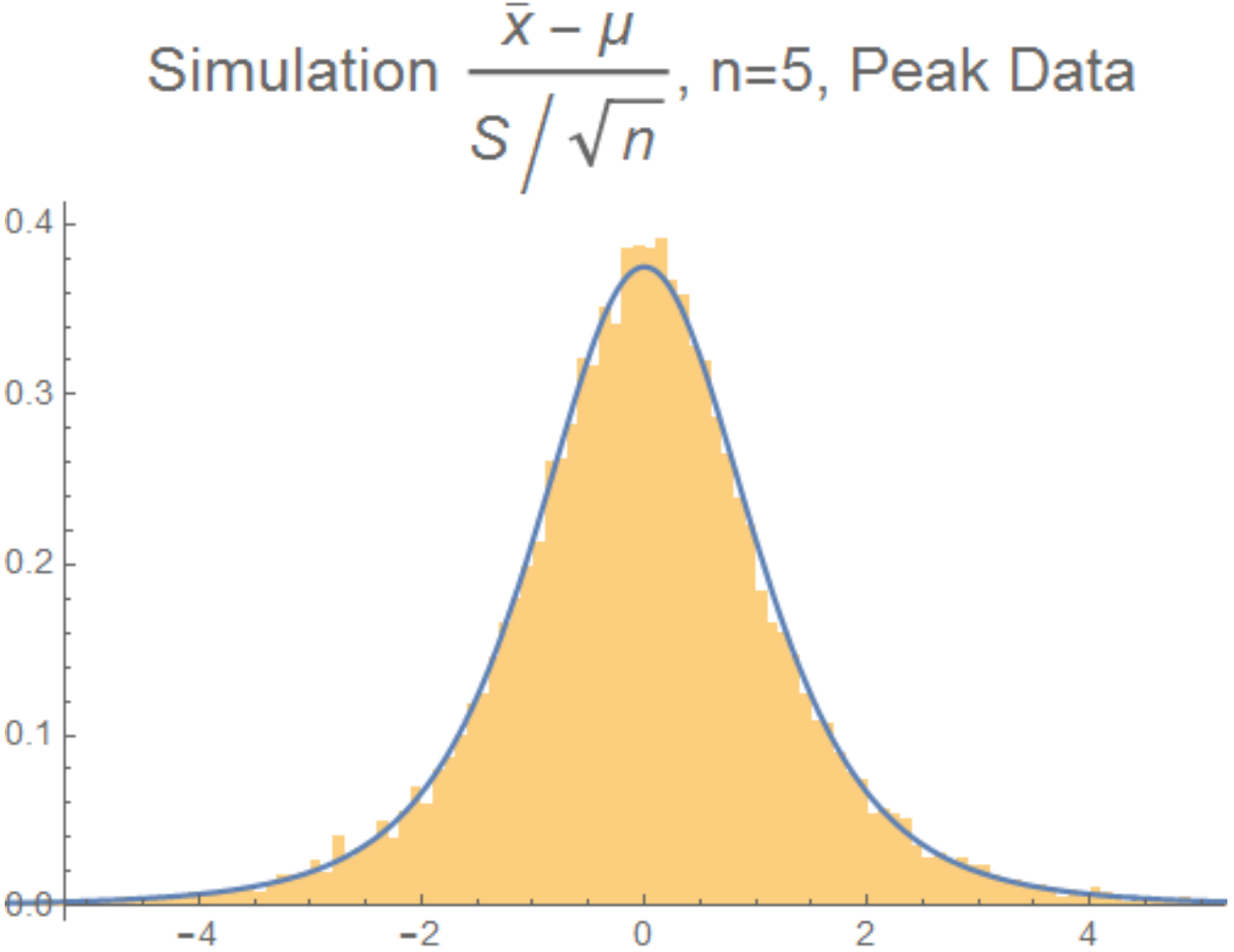
Robustness of t Tests, V

Here are simulations of the t-statistic for sampling a uniform distribution:



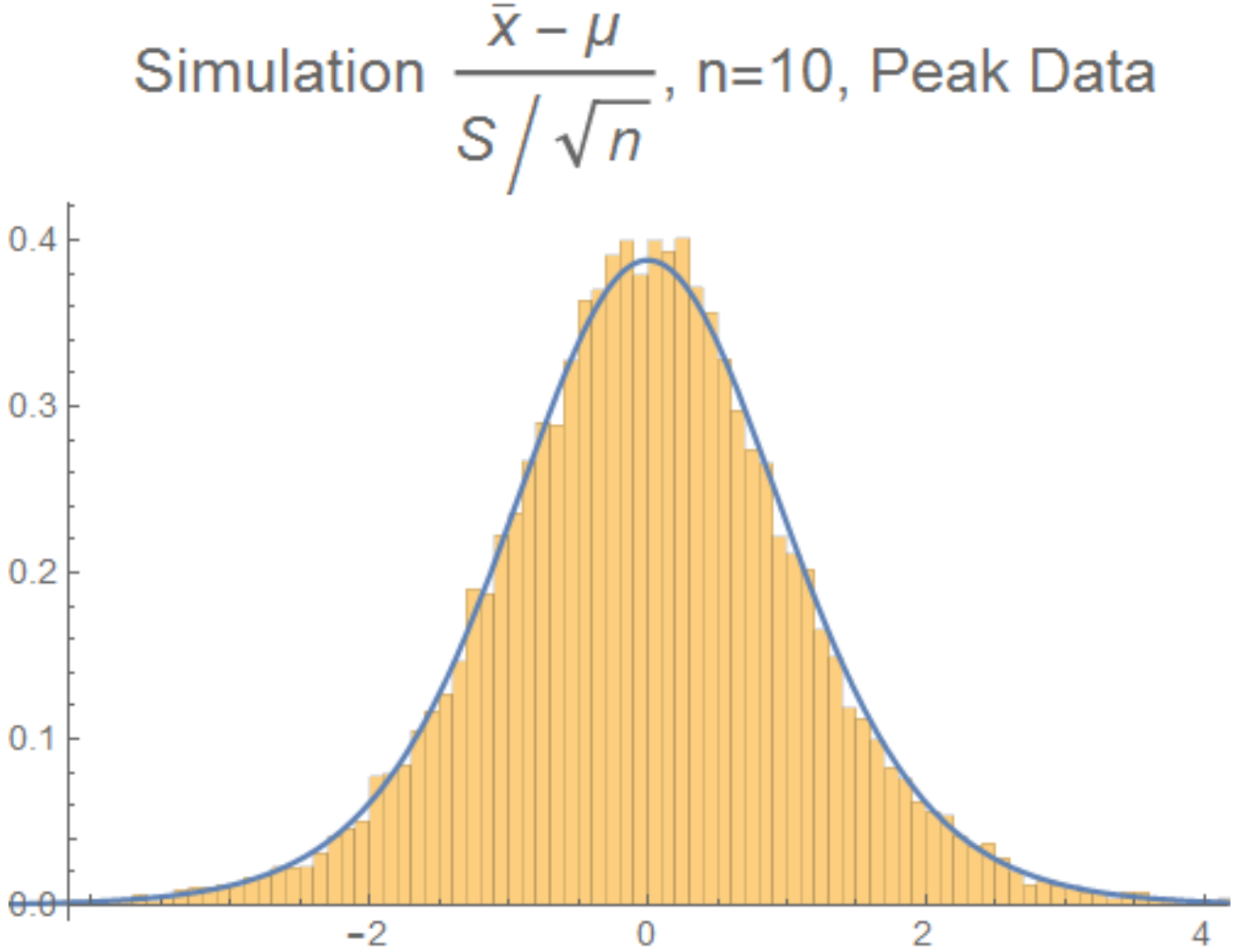
Robustness of t Tests, VI

Here are simulations of the t-statistic for sampling the "peak" distribution with $p(x) = 1 - |x|$ on $[-1, 1]$:



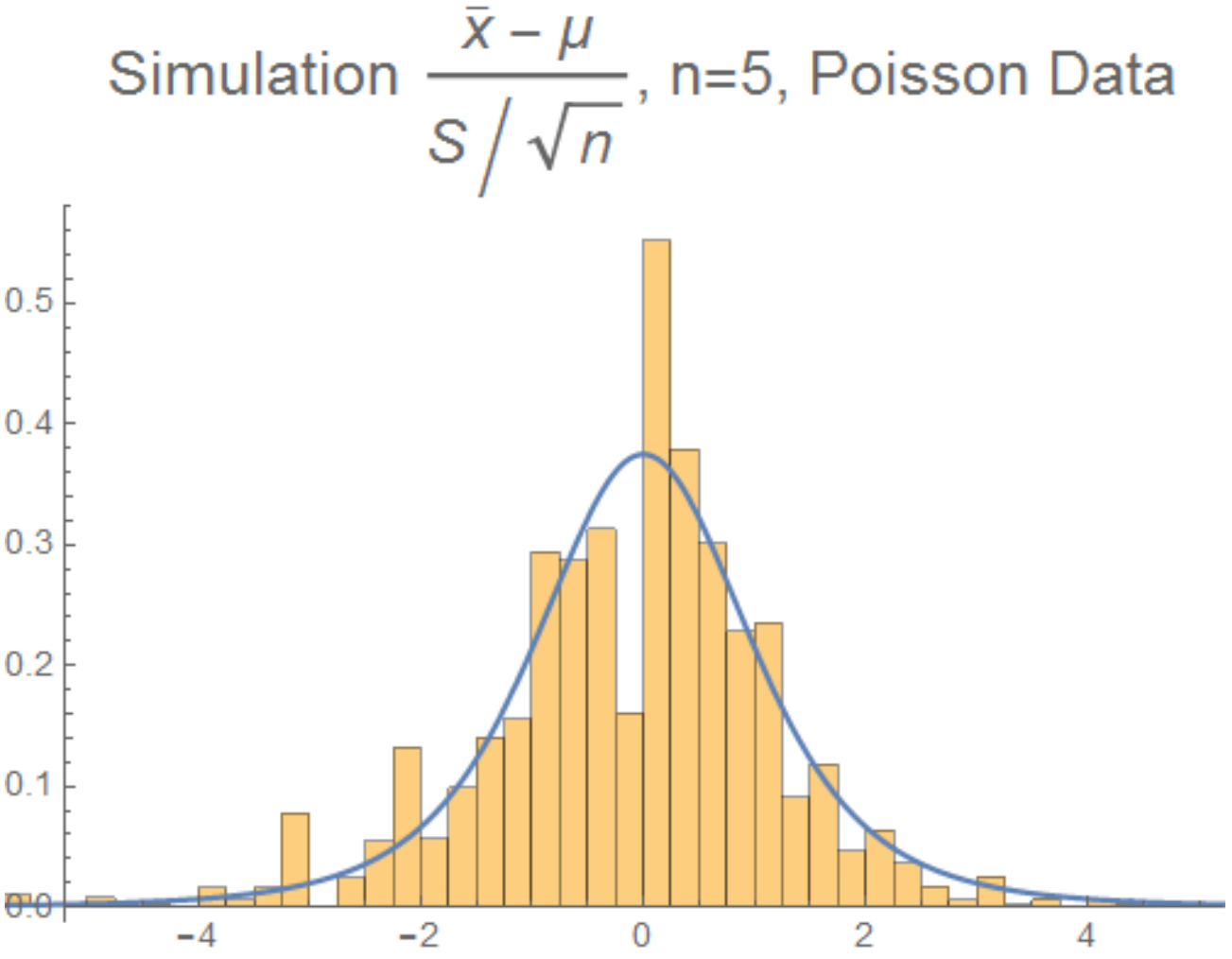
Robustness of t Tests, VII

Here are simulations of the t-statistic for sampling the "peak" distribution with $p(x) = 1 - |x|$ on $[-1, 1]$:



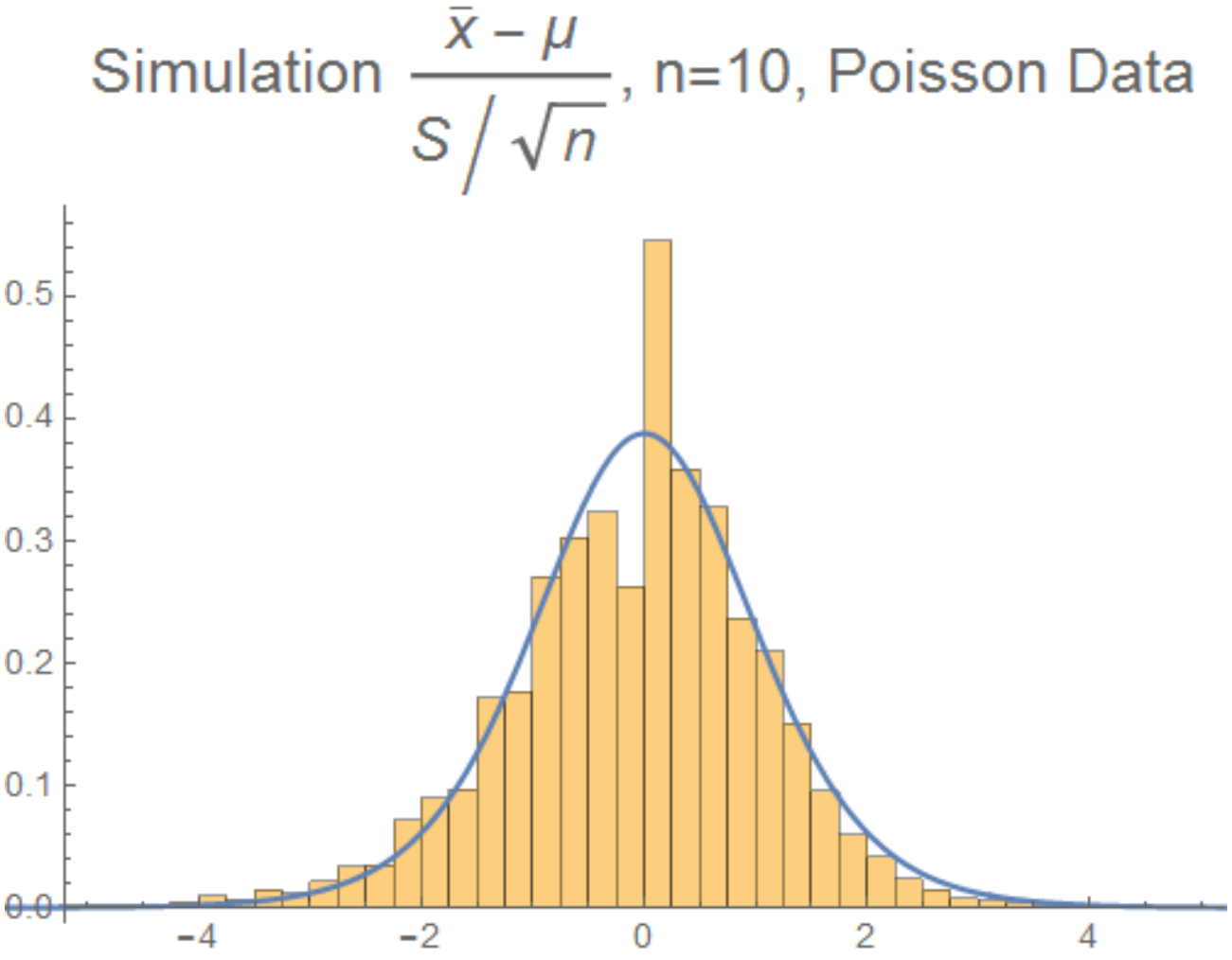
Robustness of t Tests, VIII

Here are simulations of the t-statistic for sampling the Poisson distribution with $\lambda = 3$:



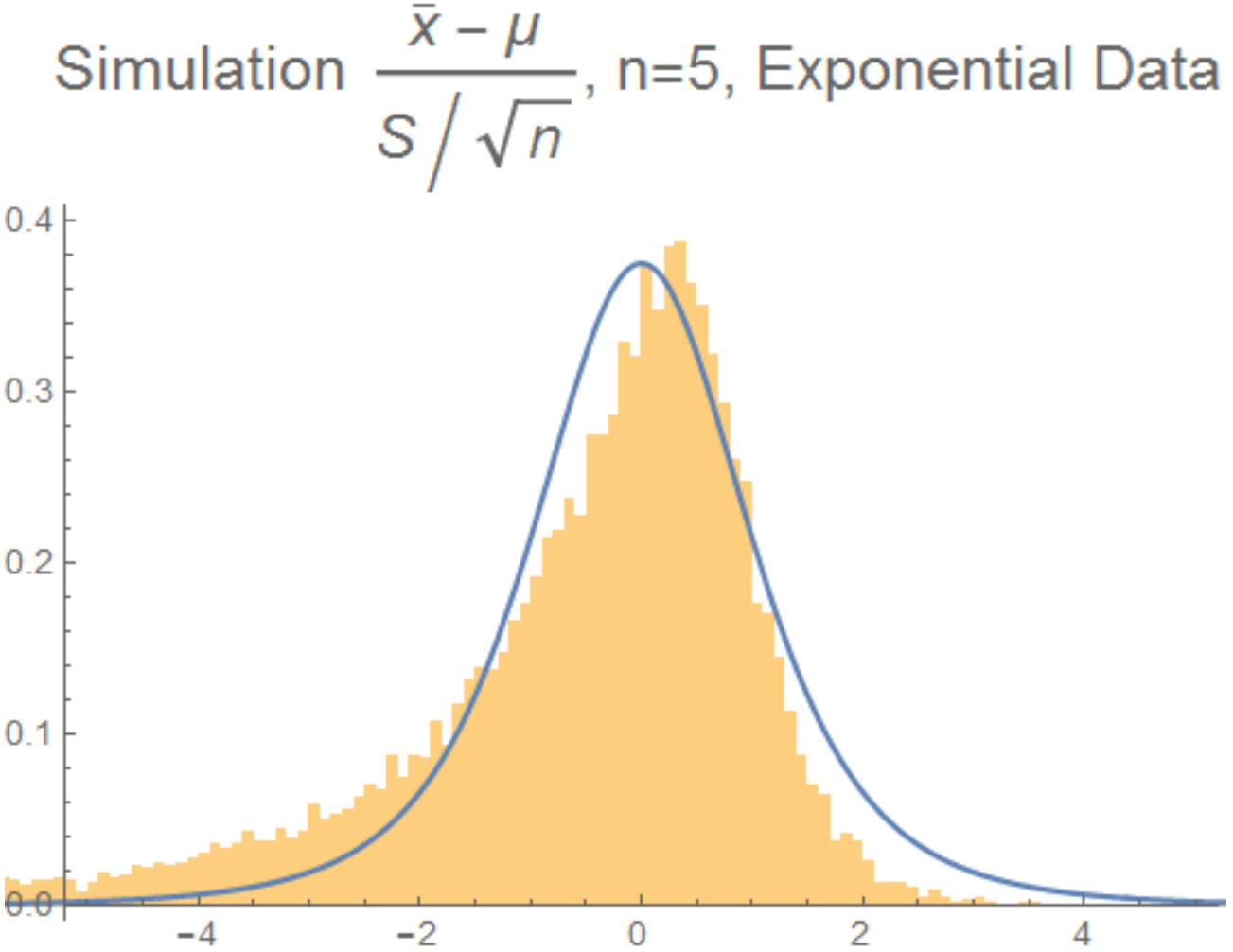
Robustness of t Tests, IX

Here are simulations of the t-statistic for sampling the Poisson distribution with $\lambda = 3$:



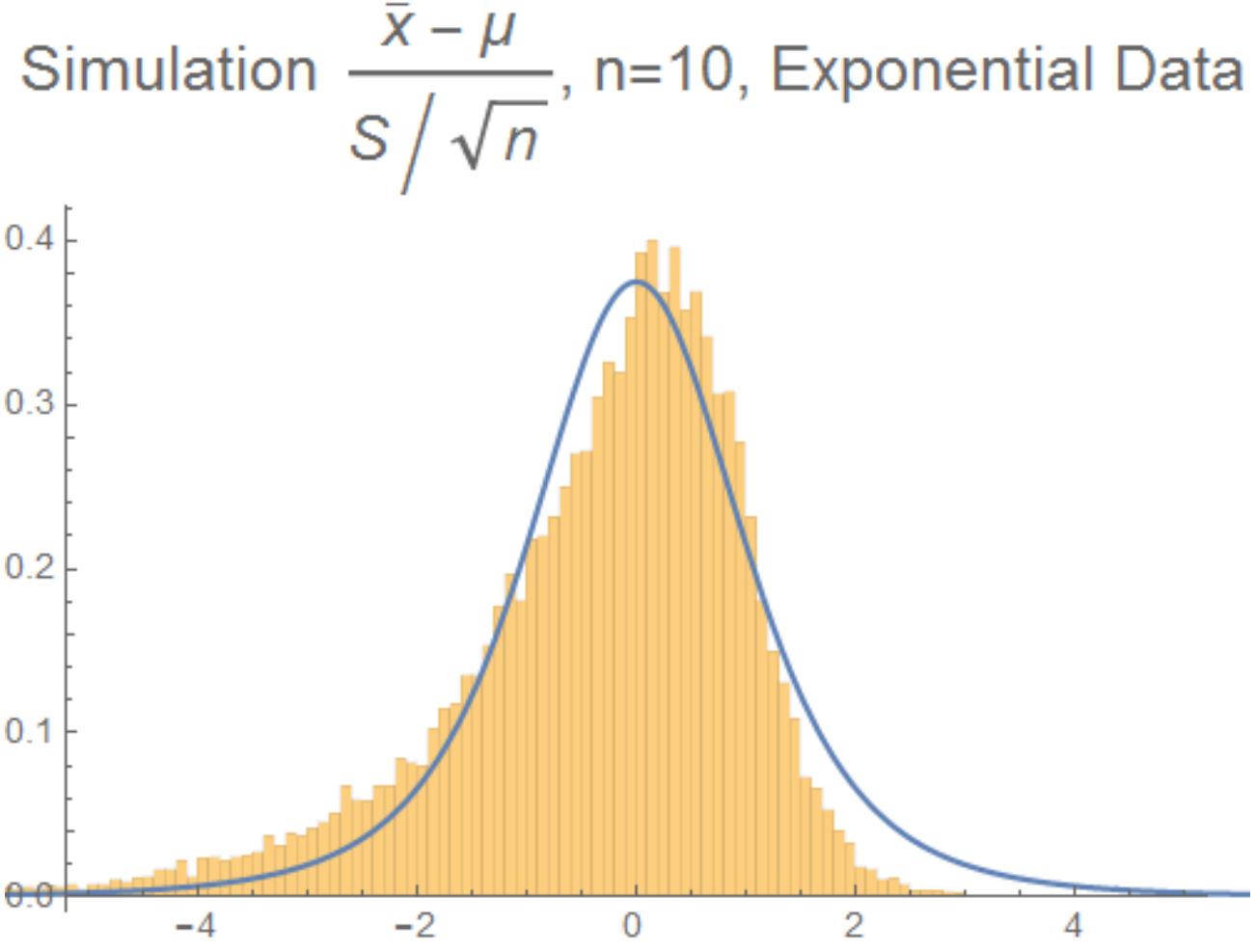
Robustness of t Tests, X

Here are simulations of the t-statistic for sampling the exponential distribution with parameter $\lambda = 1/2$:



Robustness of t Tests, XI

Here are simulations of the t-statistic for sampling the exponential distribution with parameter $\lambda = 1/2$:



Robustness of t Tests, XII

We can see from the simulations that the t distribution is fairly close for the uniform and peak distributions, it is off a bit for the Poisson, and it is very far off for the exponential.

- The uniform and peak distributions are both symmetric and do not have wide tails.
- The Poisson distribution is more skewed and has a long tail. It also has the difficulty that it is discrete and that small samples will sometimes yield all identical values (giving a sample standard deviation of 0, yielding an undefined test statistic): this explains the peculiar spike at 0.
- The exponential distribution is very skewed, which causes the resulting test statistic also to be skewed. We can see that the t distribution is not a very good model here even with a sample size $n = 10$.

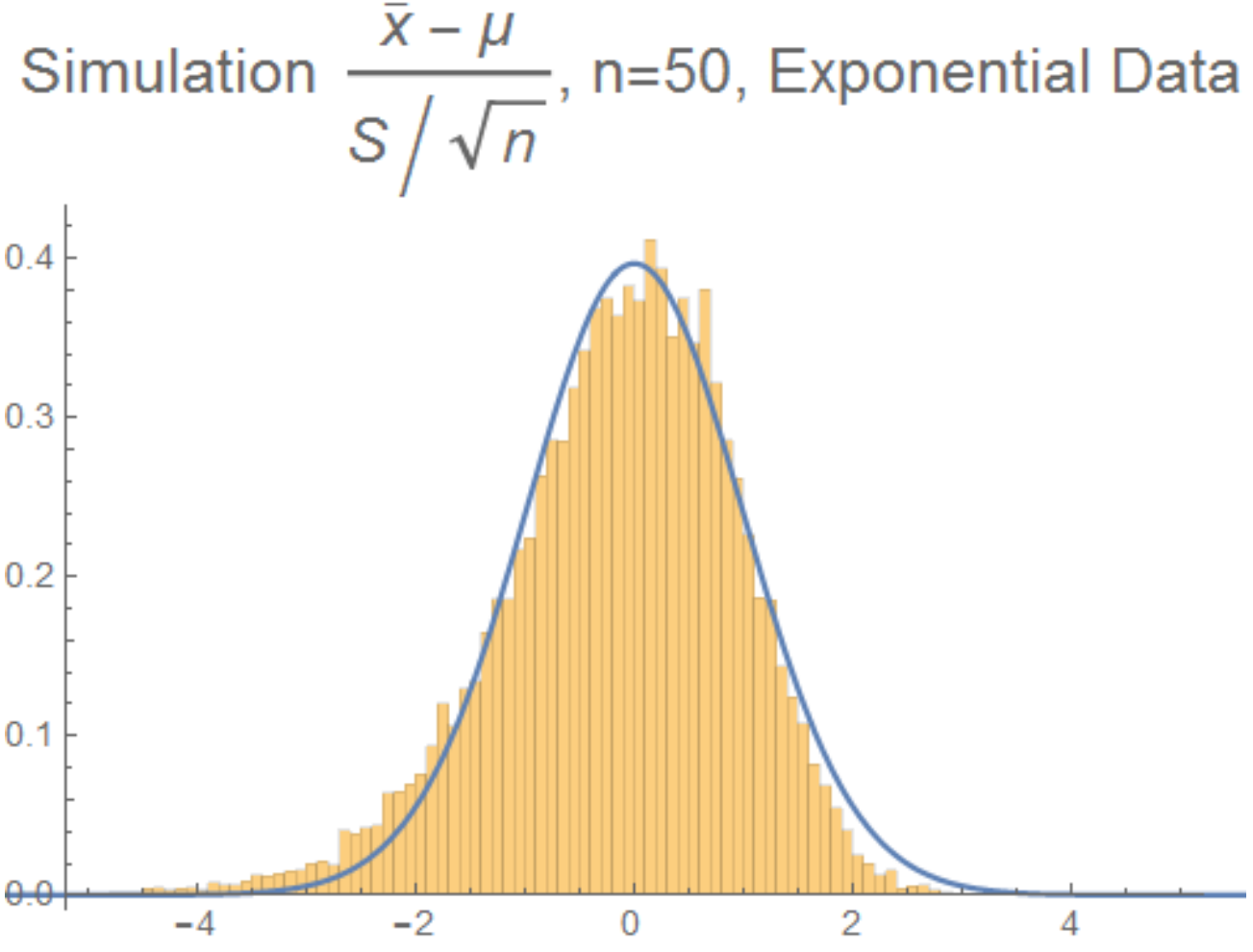
Robustness of t Tests, XIII

In general, the t distribution models the sample statistic $\frac{\overline{x} - \mu}{S/\sqrt{n}}$ well when the underlying distribution is symmetric, but not when the underlying distribution is asymmetric or skewed to one side.

- Thus, when the underlying distribution is asymmetric / skewed, t tests will not give reliable results with small samples.
- With large sample sizes (the exact definition of large, of course, depends on the scenario, but as we have seen in our discussion of the central limit theorem, usually $n = 100 - 200$ or so is quite sufficient), the central limit theorem will eventually take over and cause the sample average to be approximately normally distributed, even if the original distribution was asymmetric or skewed.
- In such cases, since the t distribution is so close to the normal distribution, either the t test or the z test will be reliable.

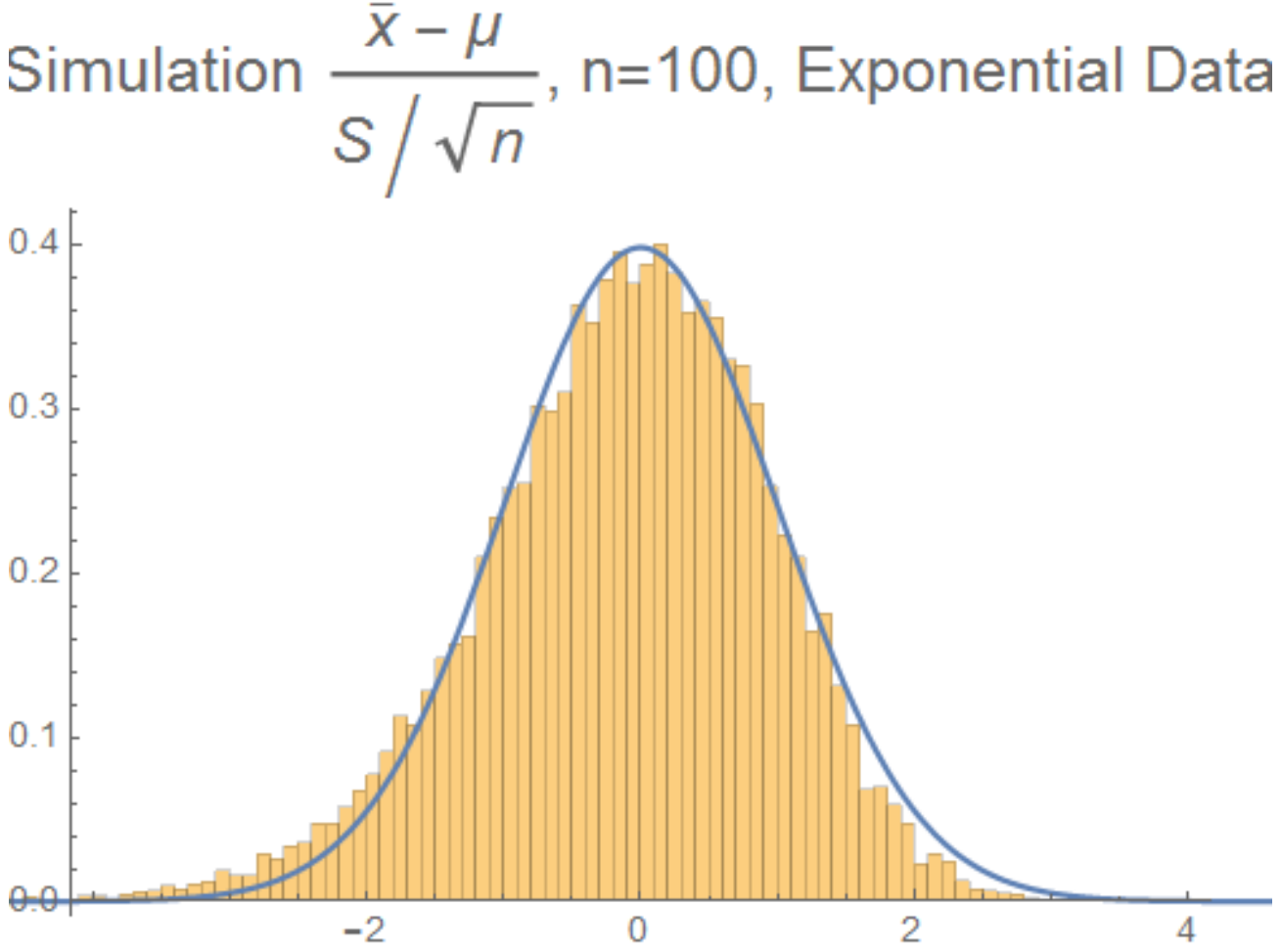
Robustness of t Tests, XIV

Here are the results of simulating the test statistic for larger n with exponentially distributed data:



Robustness of t Tests, XV

We can see here that although there is still some skewness in the histogram, it is now better approximated by the t distribution:



More t Tests, I

Example: A campus group wants to investigate whether male and female faculty are paid equally. A random sample of faculty is selected: the 50 male faculty have an average salary of \$142,081 with a sample standard deviation of \$47,683 while the 50 female faculty have an average salary of \$118,956 with a sample standard deviation of \$44,549. Assume salaries are normally distributed.

1. Use Student's equal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% CI for the difference.
2. Use Welch's unequal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% CI for the difference.

More t Tests, II

Example: The 50 male faculty have avg salary \$142,081 with std dev \$47,683, while the 50 female faculty have avg salary \$118,956 with std dev \$44,549. Assume salaries are normally distributed.

1. Use Student's equal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% CI for the difference.

More t Tests, II

Example: The 50 male faculty have avg salary \$142,081 with std dev \$47,683, while the 50 female faculty have avg salary \$118,956 with std dev \$44,549. Assume salaries are normally distributed.

1. Use Student's equal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% CI for the difference.

- Our hypotheses are $H_0 : \mu_m = \mu_f$ with alternative $H_a : \mu_m \neq \mu_f$.
- The pooled standard deviation is $S_{\text{pool}} = \$\sqrt{\frac{(50-1)\cdot 47683^2 + (50-1)\cdot 44549^2}{50+50-2}} = \$46142$, with $df = 50 + 50 - 2 = 98$.
- Our test statistic is then $t = \frac{\hat{\mu}_m - \hat{\mu}_f}{S_{\text{pool}}\sqrt{\frac{1}{n_m} + \frac{1}{n_f}}} = 2.5058$, so the p-value is $2P(T_{98} \geq 2.5058) = 0.0139$.

More t Tests, III

Example: The 50 male faculty have avg salary \$142,081 with std dev \$47,683, while the 50 female faculty have avg salary \$118,956 with std dev \$44,549. Assume faculty salaries are approximately normally distributed.

1. Use Student's equal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% confidence interval for the difference between the two groups' average salaries.

- Since the p-value is below 3%, we reject the null hypothesis and conclude that there is a difference between the two groups' average salaries.
- With $df = 98$ and $\alpha = 5\%$, we have $t_{\alpha/2,df} = 1.9845$, so the 95% CI is $(\$142081 - \$118956) - 1.9845 \cdot \$46142\sqrt{\frac{1}{50} + \frac{1}{50}} = (\$4704, \$41546)$.

More t Tests, IV

Example: The 50 male faculty have avg salary \$142,081 with std dev \$47,683, while the 50 female faculty have avg salary \$118,956 with std dev \$44,549. Assume salaries are normally distributed.

2. Use Welch's unequal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% CI for the difference.

More t Tests, IV

Example: The 50 male faculty have avg salary \$142,081 with std dev \$47,683, while the 50 female faculty have avg salary \$118,956 with std dev \$44,549. Assume salaries are normally distributed.

2. Use Welch's unequal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% CI for the difference.

- Our hypotheses are $H_0 : \mu_m = \mu_f$ with alternative $H_a : \mu_m \neq \mu_f$.
- The unpooled standard deviation is $S_{\text{unpool}} = \$\sqrt{\frac{47683^2}{49} + \frac{44549^2}{49}} = \$9228$, with $df = 97.55$.
- Our test statistic is then $t = \frac{\hat{\mu}_m - \hat{\mu}_f}{S_{\text{unpool}}} = 2.5058$, so the p-value is $2P(T_{97.55} \geq 2.5058) = 0.0139$.

More t Tests, V

Example: The 50 male faculty have avg salary \$142,081 with std dev \$47,683, while the 50 female faculty have avg salary \$118,956 with std dev \$44,549. Assume faculty salaries are approximately normally distributed.

2. Use Welch's equal-variances t-test to test at the 3% level whether the two groups have different average salaries, and give a 95% confidence interval for the difference between the two groups' average salaries.

- Since the p-value is below 3%, we reject the null hypothesis and conclude that there is a difference between the two groups' average salaries.
- With $df = 97.55$ and $\alpha = 5\%$, we have $t_{\alpha/2,df} = 1.9846$, so the 95% CI is $(\$142081 - \$118956) - 1.9846 \cdot \$9228$ $= (\$4811, \$41439)$.

More t Tests, VI

Example: The campus group brings these results to the university administration, who agree that it compellingly indicates that male faculty (avg salary \$142,081) are paid more on average than female faculty (avg salary \$118,956). The administration wants to know which division is responsible, so they break the data down further:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

3. Test at the 3% level whether average salaries for male STEM faculty are higher or lower than for female STEM faculty.
4. Test at the 3% level whether average salaries for male HSS faculty are higher or lower than for female HSS faculty.
5. Briefly describe what conclusions one may draw about the original question of whether female faculty are underpaid.

More t Tests, VII

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

3. Test at the 3% level whether average salaries for male STEM faculty are higher or lower than for female STEM faculty.

More t Tests, VII

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

3. Test at the 3% level whether average salaries for male STEM faculty are higher or lower than for female STEM faculty.

- For STEM, $H_0 : \mu_m = \mu_f$ with alternative $H_a : \mu_m < \mu_f$.
- If we use Student's t-test, the pooled standard deviation is $S_{\text{pool}} = \$\sqrt{\frac{(40-1)\cdot 30587^2+(10-1)\cdot 32578^2}{40+10-2}} = \$30970$, with $df = 40 + 10 - 2 = 48$.
- Our test statistic is $t = \frac{\hat{\mu}_m - \hat{\mu}_f}{S_{\text{pool}}\sqrt{\frac{1}{n_m}+\frac{1}{n_f}}} = -3.0551$, so the p-value is $P(T_{48} \leq -3.0551) = 0.00183$ for Student's t-test.

More t Tests, VIII

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

3. Test at the 3% level whether average salaries for male STEM faculty are higher or lower than for female STEM faculty.

More t Tests, VIII

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

3. Test at the 3% level whether average salaries for male STEM faculty are higher or lower than for female STEM faculty.

- If instead we use Welch's t-test, the unpooled standard deviation is $S_{\text{unpool}} = \$11381$ with $df = 13.26$.
- Our test statistic is $t = \dfrac{\hat{\mu}_m - \hat{\mu}_f}{S_{\text{pool}}\sqrt{\frac{1}{n_m}+\frac{1}{n_f}}} = -2.9393$ and the p-value is $P(T_{13.26} \leq -2.9393) = 0.00565$ for Welch's t-test.
- In either case, the p-value is below 3%, so we reject the null hypothesis. Our interpretation is that for STEM faculty, the female average salary is higher.

More t Tests, IX

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

4. Test at the 3% level whether average salaries for male HSS faculty are higher or lower than for female HSS faculty.

More t Tests, IX

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

4. Test at the 3% level whether average salaries for male HSS faculty are higher or lower than for female HSS faculty.

- For HSS, $H_0 : \mu_m = \mu_f$ with alternative $H_a : \mu_m < \mu_f$.
- If we use Student's t-test, the pooled standard deviation is $S_{\text{pool}} = \$21738$, with $df = 40 + 10 - 2 = 48$.
- Our test statistic is then $t = \dfrac{\hat{\mu}_m - \hat{\mu}_f}{S_{\text{pool}}\sqrt{\frac{1}{n_m} + \frac{1}{n_f}}} = -4.2317$, so the p-value is $P(T_{48} \leq -4.2317) = 0.000052$.

More t Tests, X

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

4. Test at the 3% level whether average salaries for male HSS faculty are higher or lower than for female HSS faculty.

More t Tests, X

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

4. Test at the 3% level whether average salaries for male HSS faculty are higher or lower than for female HSS faculty.

- If instead we use Welch's t-test, the unpooled standard deviation is $S_{\text{unpool}} = \$8585$ with $df = 12.32$.
- Our test statistic is $t = \dfrac{\hat{\mu}_m - \hat{\mu}_f}{S_{\text{pool}}\sqrt{\frac{1}{n_m}+\frac{1}{n_f}}} = -3.7884$ and the p-value is $P(T_{12.32} \leq -3.7884) = 0.000124$ for Welch's t-test.
- In either case, the p-value is below 3%, so we reject the null hypothesis. Our interpretation is that for HSS faculty, the female average salary is higher.

More t Tests, XI

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

5. Briefly describe what conclusions one may draw about the original question of whether female faculty are underpaid.

More t Tests, XI

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

5. Briefly describe what conclusions one may draw about the original question of whether female faculty are underpaid.

- This is where things become tricky, because if we look at both divisions together, we saw very clearly that the average salary for female faculty was less than for male faculty (statistically significant at the 3% level).
- On the other hand, in the STEM division, the average salary for female faculty was greater than that for male faculty (significant at the 3% level), and the same also held for the HSS division!

More t Tests, XII

Example: Salary data broken down by division:
STEM faculty: 40 male (avg \$160691, std dev \$30587).
10 female (avg \$194143, std dev \$32578).
HSS faculty: 10 male (avg \$67638, std dev \$25056)
40 female (avg \$100160, std dev \$20897).

5. Briefly describe what conclusions one may draw about the original question of whether female faculty are underpaid.

- It seems hard to reconcile those three facts: female faculty are paid more in each individual division, but less overall.
- Here's why: there are more male faculty in STEM, which also has higher salaries than HSS. So, even though the female faculty are better paid in each division, more of them are in the lower-paying division, so their overall average is lower.
- This is an example of Simpson's paradox: it is the phenomenon where a trend or result holds inside of subgroups, but is reversed when the groups are combined.

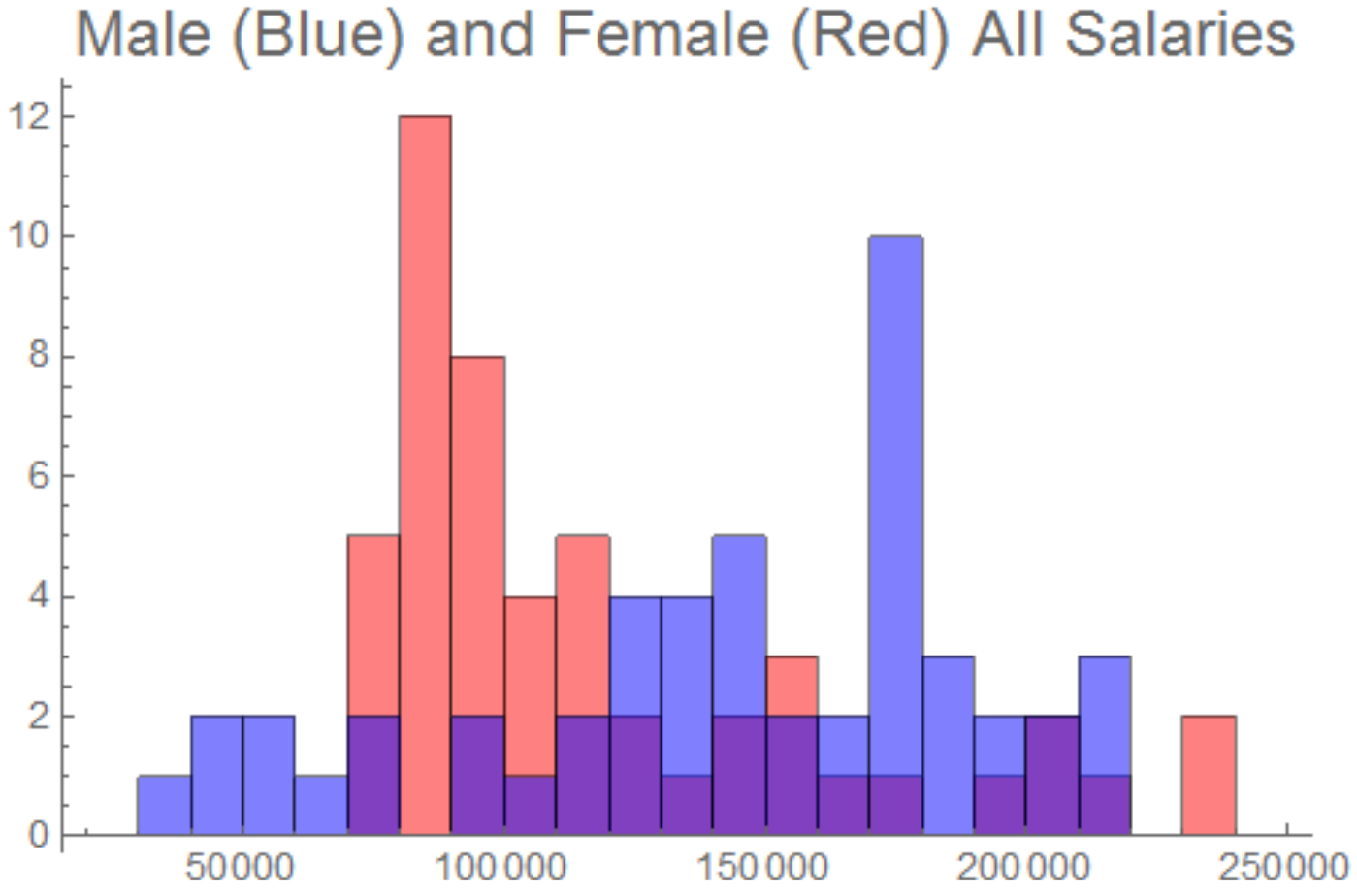
More t Tests, XIV

Simpson's paradox presents a serious problem when using statistics and grouping populations together.

- These statistical methods alone are not capable of identifying whether there is a "confounding" (or "lurking") variable that may be affecting the results but is not included in the collected data.
- Simpson's paradox also yields difficult-to-reconcile conclusions, such as in this case with faculty salaries.
- It is very easy to give misleading statistics: here, both the "male faculty are paid more" and "female faculty are paid more" camps have t-tests showing that they're "correct"!
- This is one more reason why it is so important to study an experimental question from many different directions, and use a variety of statistical summaries.

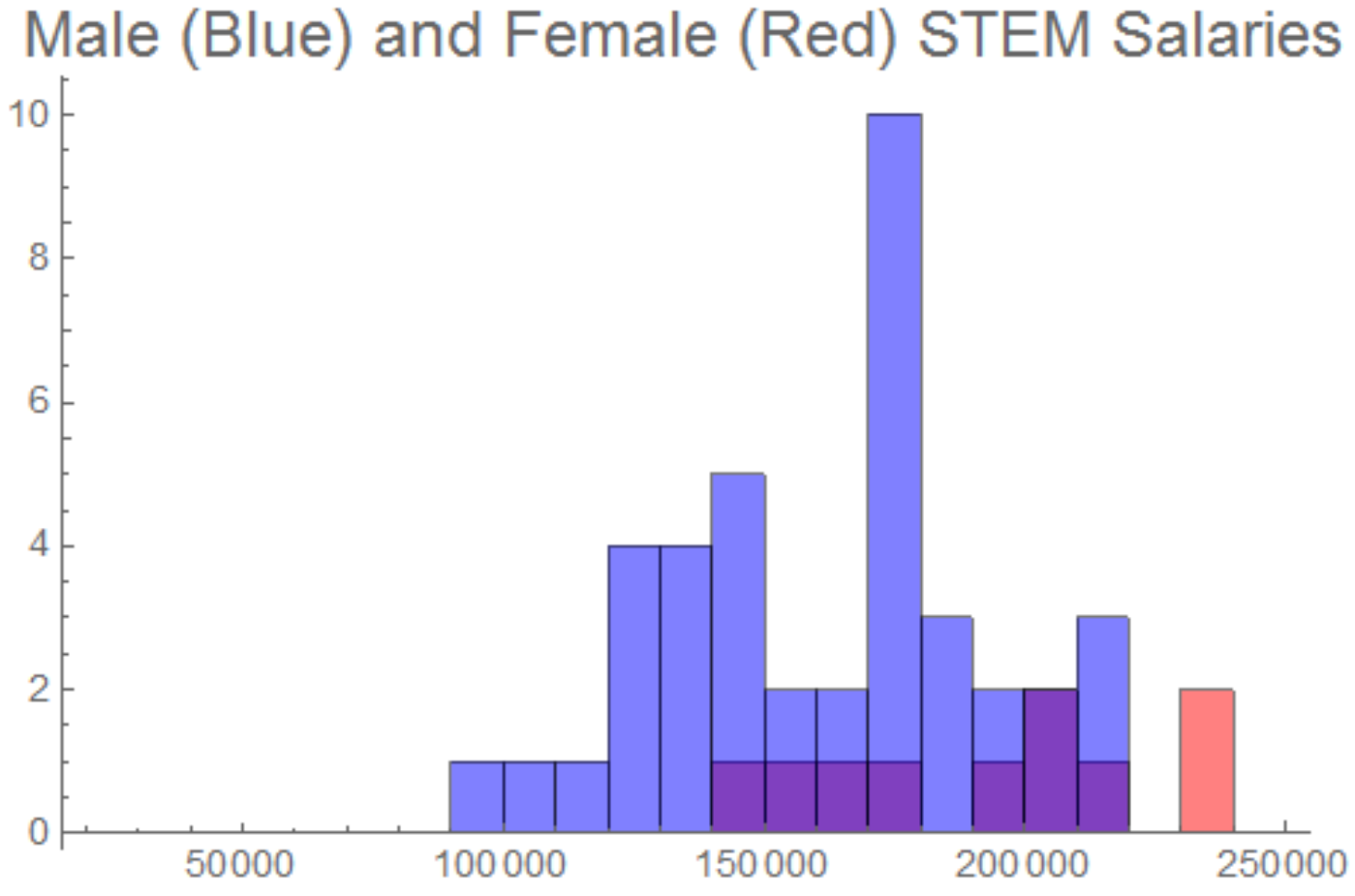
More t Tests, XV

Looking at the actual distributions might have suggested that the presence of another variable that might be relevant (or at the very least, that the overall distributions look quite skewed):



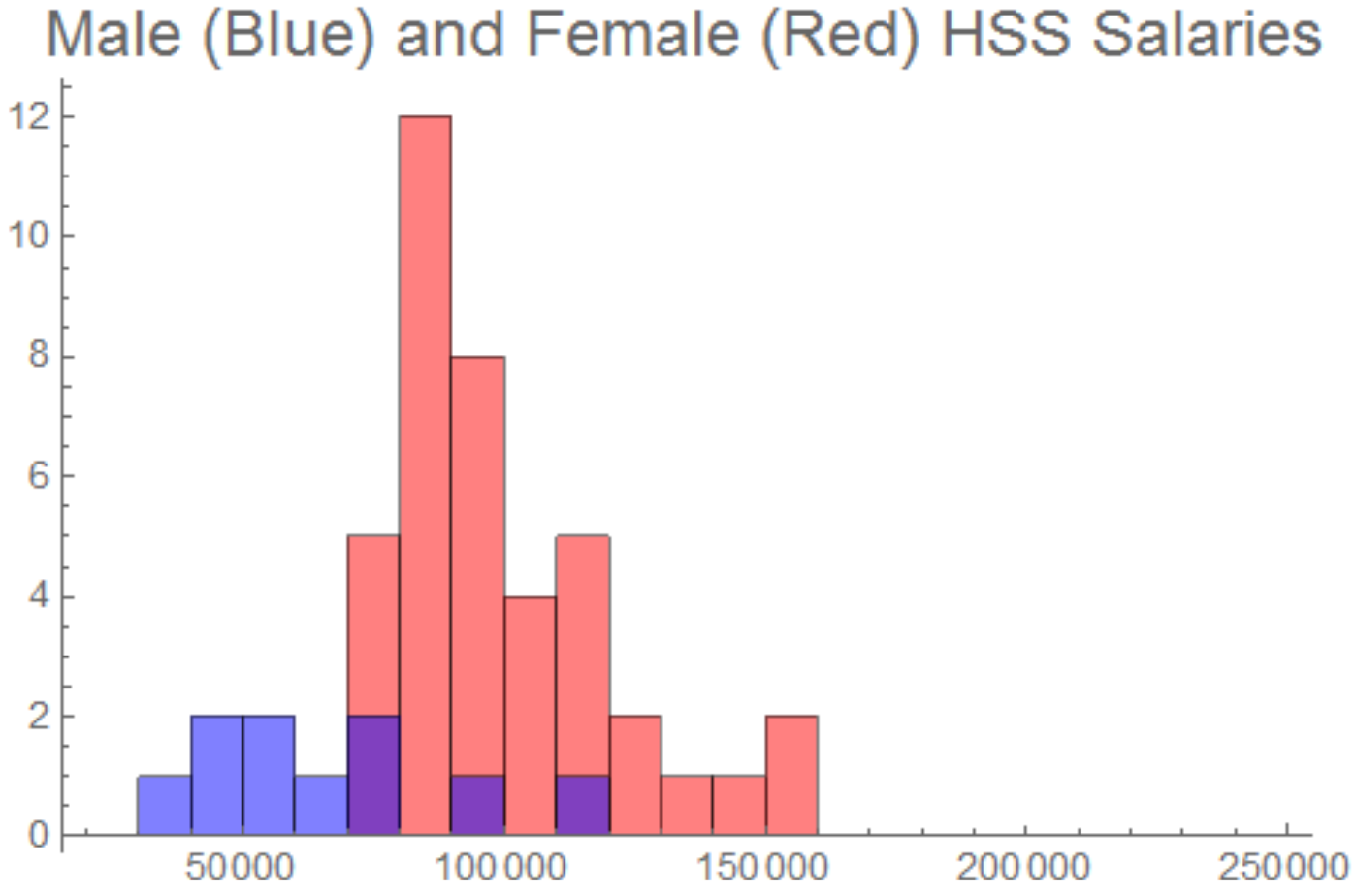
More t Tests, XVI

Looking at the actual distributions might have suggested that the presence of another variable that might be relevant (or at the very least, that the overall distributions look quite skewed):



More t Tests, XVII

Looking at the actual distributions might have suggested that the presence of another variable that might be relevant (or at the very least, that the overall distributions look quite skewed):



Summary

We discussed matched-pairs tests.

We discussed the robustness of t tests.

We did more examples of t tests.

Next lecture: The χ^2 distribution and estimating variance