## Sex and Salary: Example of Analysis with Independent Samples t Test

One of my graduate students found, online, an SPSS data set with, among other variables, data on employees' sex and salaries. No description of the source of the data is available online.

As is the usual case with researchers in Psychology, as soon as I got my hands on the data I jumped into an analysis without first checking the assumptions of the analysis. I conducted a $t$ test comparing the salaries of women with those of men. Here are selected parts of the output, with comments.

T-TEST GROUPS=Sex(10)
/MISSING=ANALYSIS
/VARIABLES=salary
/CRITERIA=CI(.95)

Group Statistics

|  | Sex |  |  |  | Std. Error <br> Mean |
| :--- | :--- | ---: | ---: | ---: | ---: |
| salary | Female | 216 | 26031.92 | Std. Deviation | 7558.021 |
|  | Male | 258 | 41441.78 | 19499.214 | 1214.258 |

The ratio of the two variances here is 6.66 , so we shall use the separate variances test.

|  |  | Independent Samples Test |  |  |  |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | t | df | Sig. (2-tailed) | quality of Mea |  |  |
|  |  | Mean <br> Difference |  |  | 95\% Confidence Interval of the Difference |  |
|  |  | Lower |  |  | Upper |
| salary | Equal variances assumed |  | 10.945 | 472 | . 000 | 15409.862 | 12643.322 | 18176.401 |
|  | Equal variances not assumed | 11.688 | 344.262 | . 000 | 15409.862 | 12816.728 | 18002.996 |

Holy moly, the men were making, on average, $\$ 15,410$ more per year than were the women. With a $95 \%$ confidence interval running from $\$ 12,817$ to $\$ 18,003$. From the means here, we can tell these data are dated, from a time where a dollar was worth more than what it is now. The unit of measure here is intrinsically meaningful for those who know the value of a dollar, but I estimated Cohen's $\delta$ nevertheless: $d=1.10,95 \% \mathrm{Cl}[.82,1.20]$. That is a large effect.

The data set includes data on seniority - month on the job. Maybe the men got the higher salaries because they had greater seniority. I'll add seniority to the model as a covariate.

```
UNIANOVA salary BY Sex WITH jobtime
    /METHOD=SSTYPE(3)
    /INTERCEPT=INCLUDE
    /EMMEANS=TABLES(Sex) WITH(jobtime=MEAN)
    /CRITERIA=ALPHA(.05)
```

Between-Subjects Factors

|  |  | Value Label | N |
| :--- | :--- | :--- | :--- |
| Sex | .00 | Female | 216 |
|  | 1.00 | Male | 258 |

Tests of Between-Subjects Effects
Dependent Variable: salary

| Source | Type III Sum of Squares | df | Mean Square | F | Sig. |
| :--- | ---: | ---: | ---: | ---: | ---: |
| Corrected Model | $28325290000.000^{\mathrm{a}}$ | 2 | 14162645000.000 | 60.868 | .000 |
| Intercept | 4923876057.000 | 1 | 4923876057.000 | 21.162 | .000 |
| jobtime | 406756970.900 | 1 | 406756970.900 | 1.748 | .187 |
| Sex | 27350012210.000 | 1 | 27350012210.000 | 117.545 | .000 |
| Error | 109591205400.000 | 471 | 232677718.500 |  |  |
| Total | 699467436900.000 | 474 |  |  |  |
| Corrected Total | 137916495400.000 | 473 |  |  |  |

a. R Squared $=.205$ (Adjusted R Squared $=.202$ )

The Analysis of Covariance shows that even if men and women had equal levels of seniority, the men would have significantly higher salaries. How much higher?

## Estimated Marginal Means

|  |  | Sex |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Depend | ariable: |  |  |  |
|  |  |  | 95\% Confide | nce Interval |
| Sex | Mean | Std. Error | Lower Bound | Upper Bound |
| Female | $26099.363^{\text {a }}$ | 1039.141 | 24057.437 | 28141.288 |
| Male | $41385.320^{\text {a }}$ | 950.618 | 39517.342 | 43253.298 |

a. Covariates appearing in the model are evaluated at the following values: jobtime $=81.11$.

The estimated marginal means (aka least squares means or adjusted means) are estimates of what the group salaries would be if the groups did not differ on the covariate (seniority). The difference here is $\$ 15,286$, only a minor reduction from what it was when we ignored seniority.

## I Need a Spanking

Good statisticians know that they should evaluate the assumption of the procedure they intend to use prior to using it. I did evaluate the homogeneity of variance assumption, but not the normality assumption. Shame on me. In my experience, heterogeneity of variance is often accompanied by skewed distributions, so l'd best look at the within-group distributions of salary. I split the file by sex and got some basic descriptive stats and plots.

|  | Statistics $^{\text {a }}$ |  |
| :--- | :--- | ---: |
| salary |  |  |
| N | Valid | 216 |
|  | Missing | 0 |
| Skewness |  | 1.863 |
| Kurtosis |  | 4.641 |

a. Sex = Female

Yikes, the salaries are quite skewed among the women (I become concerned when the absolute value of skewness is one or higher). The high kurtosis indicates the presence of outlier in the data. The plot below shows that the scores have a distinct positive skewness with a cluster of outliers on the right side. The smoothed curve is normal curve with the same mean and variance as our sample.


Sex = Male

| Statistics $^{\text {a }}$ |  |  |
| :--- | :--- | ---: |
| salary |  |  |
| N | Valid | 258 |
|  | Missing | 0 |
| Skewness |  | 1.639 |
| Kurtosis |  | 2.780 |

a. Sex = Male

Same problem with the men's salaries. Fortunately the skewness among the men's salaries is not much different from that among the women's salaries, so a transformation that normalizes the
scores in the one group will likely do the same in the other group.


With positively skewed data, a square root or a log transformation often normalizes the data. I used a log transformation and rechecked the distributions.

COMPUTE Log_Salary=LG10(salary).
EXECUTE.

## Sex = Female

| Statistics $^{\text {a }}$ |  |  |
| :--- | :--- | ---: |
| Log_Salary |  |  |
| N | Valid | 216 |
|  | Missing | 0 |
| Skewness |  | .858 |
| Kurtosis |  | 1.201 |

a. Sex $=$ Female


Paranormal Distribution


## Sex = Male

Statistics ${ }^{\text {a }}$

| Log_Salary |  |  |
| :--- | :--- | ---: |
| N | Valid | 258 |
|  | Missing | 0 |
| Skewness |  | .845 |
| Kurtosis |  | -.174 |

a. Sex = Male


That looks much better. Now l'll conduct a $t$ test on the transformed data.

Group Statistics

|  |  |  |  |  | Std. Error <br> Mean |
| :--- | :--- | ---: | ---: | ---: | ---: |
| Log_Salary | Female | N | Mean | Std. Deviation | Mea |
|  | Male | 216 | 4.4005 | .11033 | .00751 |

The transformation also reduced the heterogeneity of variance. Now the variance in the men's salaries on only 2.4 times that in the women's salaries.

Independent Samples Test

|  |  | t | df | -test for Equality of Means |  |  |
| :---: | :---: | :---: | :---: | :---: | :---: | :---: |
|  |  | Sig. (2-tailed) |  | 95\% Confidence Interval of the Difference |  |
|  |  | Lower |  | Upper |
| Log_Salary | Equal variances assumed |  | 13.127 | 472 | . 000 | . 15220 | . 20579 |
|  | Equal variances not assumed | 13.629 | 442.401 | . 000 | . 15318 | . 20481 |

The difference remains significant. It would confuse readers were I report the group means in log dollars, so l'll report the original means.

As shown in Table 1, the distribution of salaries was considerably skewed in both women and men, but a base ten log transformation normalized the distributions. The mean salary in men was significantly higher in men than in women, $t(442.4)=13.629, p<.001, d=1.21,95 \% \mathrm{Cl}[1.01,1.41]$.

Table 1
Sex Differences in Salaries at the Rountree Widget Factory

|  | Annual Salary (\$) |  |  |  |
| :---: | :---: | :---: | :---: | :---: |
| Group | $M$ | $S D$ | $n$ | $s k$ |
| Women | 26,032 | 7,558 | 216 | 1.86 |
| Men | 41,442 | 19,499 | 258 | 1.64 |

