# ANALYZING DATA FROM INDEPENDENT SAMPLES

Key idea: Variances add. We now have two separate groups of data. We cannot combine them into one (as we did with paired data). The two separate groups provide two separate sources of variation. However, if we know the variability associated with each group of data (and we do), then we simply combine them to get the combined variability.

## Analyzing two proportions

We are comparing the proportion of "successes" in two different groups. Examples: the percentage of men and women who are asleep in class; the proportion of lab rats in the treatment and in the control group that get cancer; whether Stetson or Rollins graduates are more likely to get a job after graduation.

The sampling standard deviation:

Recall that the standard deviation for a single proportion is

$$sd(p) = \sqrt{\frac{p \cdot (1-p)}{n}}$$

(where $p$ is the proportion in question and $n$ is the number of data points). Hence, for comparing two proportions, we will have

$$sd(p_1 - p_2) = \sqrt{\frac{p \cdot (1-p)}{n_1} + \frac{p \cdot (1-p)}{n_2}}$$

For a confidence interval:

We know the general form of a confidence interval is

$$[\text{best guess}] \pm [\#] \cdot [\text{st dev}]$$

EXAMPLE: How big is the "gender gap" in politics? Out of 800 women surveyed, 440 (55%) say they plan to vote for the Democrat candidate, while 380 out of 800 men surveyed (47.5%) say they plan to vote for a Democrat.

SOLUTION: 95% confidence interval

$$[.55 - .475] \pm [1.96] \cdot \sqrt{\frac{(.55)(1-.55)}{800} + \frac{(.475)(1-.475)}{800}}$$

$$0.075 \pm 0.049$$

NOTE that the two sample proportions used in the standard deviation formula are different (0.55 and 0.475) – there is no reason to believe they are the same.

For a hypothesis test:

We know the general form for a test statistic is

$$test\ statistic = \frac{\text{obs - exp}}{\text{sd}}$$

EXAMPLE: Are men or women more likely to fall asleep in statistics class? Dr. Rasp has historical data for several years of statistics classes. He notes that, of 180 men who have taken the class in recent years, 72 (40%) have fallen asleep in class at least once. Among the 120 women who have taken the class during the same time frame, 30 (25%) have fallen asleep at least once.

SOLUTION: This is a hypothesis test. The hypotheses are

$H_0$: men and women fall asleep equally often $\quad \pi_M = \pi_W$

$H_A$: men and women don't fall asleep equally often $\quad \pi_M \neq \pi_W$

NOTE that I have made this a two-tailed test; I had no *a priori* reason to believe men or women were more sleep-inclined. If you did have reason to suspect such a difference before you looked at the data, then you could do a one-tailed test.

NOTE one key difference between this example and the previous one: now, we are assuming that the two true proportions (sleepers for men and women) are the same. We did not make this assumption for the confidence interval. But because we assume the null is true until proven otherwise, we need to use the same estimate for the proportion of male and female sleepers. We note that overall, 102 people (72 men + 30 women) were asleep, out of 300 total (180 men + 120 women). Hence, it appears that 102/300 = 34% are sleepers. We will use this number in our standard deviation calculations.

$$z = \frac{(0.40 - 0.25) - 0}{\sqrt{\frac{(0.34)(1-0.34)}{180} + \frac{(0.34)(1-0.34)}{120}}} = 2.69$$

The Excel function `=NORM.S.DIST(2.69, TRUE)` gives a value of 0.9964. So the area in the upper tail is 1 – 0.9964 = 0.0036. Doubling this (if you are doing a two-tailed test) gives a p-value of 0.0072. This is small – we reject the null hypothesis, and conclude that men and women don't fall asleep equally often.

## Analyzing two means

We are comparing the average numerical response for two different groups. Examples: the amount of time men and women sleep in class; the survival time of lab rats in the treatment and in the control groups; starting salaries for Stetson and Rollins graduates.

The sampling standard deviation:

Recall that the standard deviation for a single mean is

$$sd(\overline{X}) = \frac{\sigma}{\sqrt{n}}$$

(where σ is the true proportion standard deviation and *n* is the number of data points). We typically do not know the true population standard deviation, and hence substitute the sample value. This results in a sampling distribution that follows Student's *t*, rather than the normal distribution.

For two means we will have

$$sd(\overline{X}_1 - \overline{X}_2) = \sqrt{\frac{Var(X)}{n_1} + \frac{Var(X)}{n_2}}$$

NOTE that we are now combining two Student's *t* distributions rather than two normal distributions. This introduces one additional, minor, complication: the two underlying variances for the two groups need to be the same.[1]

The pooled variance:

The pooled variance is the weighted average of the two sample variances. Weighting is done by degrees of freedom, rather than by sample size. In formula terms:

$$s_p^{\,2} = \frac{(n_1 - 1)\cdot Var1 + (n_2 - 1)\cdot Var2}{(n_1 - 1) + (n_2 - 1)}$$

EXAMPLE: Our study compares the amount of sleep that men and women get the night before a "knowledge festival." The eight men in the class averaged five hours of sleep, with a standard deviation of two hours (hence a variance of 4). The twelve women averaged seven hours of sleep, with a standard deviation of 1.732 hours (hence a variance of 3).

SOLUTION: We want a weighted average of the variance for men (4) and for women (3). There are 8 and 12, respectively, in the two groups. We weight by degrees of freedom, rather than sample size … which means we are averaging seven 4's and eleven 3's. Thus:

$$s_p^{\,2} = \frac{(7)(4) + (11)(3)}{7 + 11} = 3.389$$

For a confidence interval:

We know the general form of a confidence interval is

$$[\,\text{best guess}\,] \pm [\,\#\,] \cdot [\,\text{st dev}\,]$$

[1] What if they're not? Excellent question! Unfortunately, it's a bit too advanced for an introductory course. We'll go into it in more detail when you take the elective course, STAT 460 (Experimental Design).

EXAMPLE: How much more sleep, on average, do women get than men get, the night before a "knowledge festival"? (Using the data given previously.)

SOLUTION: 95% confidence interval; Student's *t* distribution

$$[7-5] \pm [2.101] \cdot \sqrt{\frac{3.389}{12} + \frac{3.389}{8}}$$

$$2 \pm 1.76$$

For a hypothesis test:

We know the general form for a test statistic is

$$\textit{test statistic} = \frac{\text{obs - exp}}{\text{sd}}$$

EXAMPLE: Do men and women differ in the amount of sleep they get, the night before a "knowledge festival"? (Using the data given previously.)

SOLUTION: The hypotheses are

$H_0$: men and women sleep equally much before a k.f. $\mu_M = \mu_W$

$H_A$: they don't $\mu_M \neq \mu_W$

NOTE that I have made this a two-tailed test; I had no *a priori* reason to believe men or women were more sleep-inclined. If you did have reason to suspect such a difference before you looked at the data, then you could do a one-tailed test.

The test statistic is

$$t = \frac{(7-5)-0}{\sqrt{\frac{3.389}{12} + \frac{3.389}{8}}} = 2.38$$

This follows Student's *t* distribution with 18 degrees of freedom. We get the p-value from Microsoft Excel. The cell formula

```
=TDIST(2.38,18,2)          (in Excel 2007 and earlier)
=T.DIST.2T(2.38,18)        (in Excel 2010 and later)
```

tells us that the p-value is 0.0286. We're going to reject the null hypothesis. There is sufficient evidence to believe that men and women differ in their average amount of sleep, the night before a "knowledge festival."