# ST 516: Foundations of Data Analytics

## Inference for means in two samples

# Motivating example

Do newborns from mothers who smoke have a different average birth weight than newborns from mothers who don't smoke?

How big is the difference?

A random sample of 150 cases of mothers and their newborns is obtained from North Carolina.

The smoking group includes 50 cases and the nonsmoking group contains 100 cases.

Four observations from this data set are presented below. We are particularly interested in two variables: *weight* and *smoke*. The *weight* variable represents the weights of the newborns and the *smoke* variable describes which mothers smoked during pregnancy.

| | fAge | mAge | weeks | weight | sexBaby | smoke |
|---|---|---|---|---|---|---|
| 1 | NA | 13 | 37 | 5.00 | female | nonsmoker |
| 2 | NA | 14 | 36 | 5.88 | female | nonsmoker |
| 3 | 19 | 15 | 41 | 8.13 | male | smoker |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | |
| 150 | 45 | 50 | 36 | 9.25 | female | nonsmoker |

# Notation for populations

There are now two population distributions:

- weights of babies born to mothers who smoke, and
- weights of babies born to mothers who don't smoke.

In general, we will refer to them as population 1 and population 2.

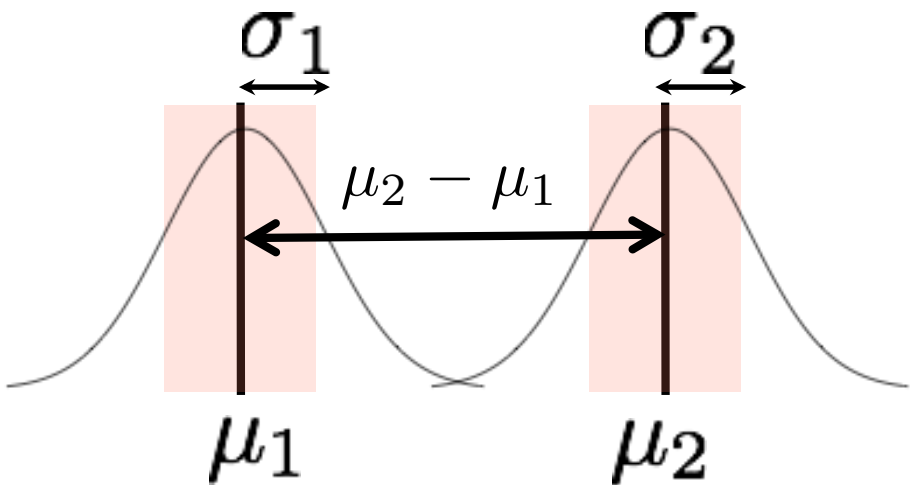
Each population distribution has a mean and standard deviation. We'll use subscripts to identify which population we are talking about:

- $\mu_{\text{smoker}}$, $\sigma_{\text{smoker}}$, or more generally $\mu_1$ and $\sigma_1$,
- $\mu_{\text{nonsmoker}}$, $\sigma_{\text{nonsmoker}}$, or more generally $\mu_2$ and $\sigma_2$,

# Notation for populations

Our questions of interest can be rephrased as questions about these population parameters:

- Is $\mu_{\text{smoker}} = \mu_{\text{nonsmoker}}$? Equivalently, is $\mu_{\text{smoker}} - \mu_{\text{nonsmoker}} = 0$?
- How big is $\mu_{\text{nonsmoker}} - \mu_{\text{smoker}}$?




population 1 with unknown mean, $\mu_1$ and unknown standard deviation $\sigma_1$

population 2 with unknown mean, $\mu_2$ and unknown standard deviation $\sigma_2$

# Notation for samples

We have two independent samples, one from each population.

To denote the sampled values we will use $X_{ij}$ where $i$ indexes the population, and $j$ the observations number. For example, $X_{12}$ is the second observations from the first population.

Our two samples can be denoted:

- $X_{11}, \ldots, X_{1n_1}$, randomly sampled from population 1
- $X_{21}, \ldots, X_{2n_2}$, randomly sampled from population 2

Notice the two samples may not have the same sample size. In our example, $n_1 = 50$ and $n_2 = 100$.

Each sample has its own sample statistics, $\overline{X}_1, \overline{X}_2, s_1, s_2$ where again the subscript denotes the population.

# Point estimate for difference in means

How might we estimate the difference in population means?

Just like in a single sample, point estimates for each population mean are the sample means.

The point estimate of the difference in population means is the difference in sample means:

$$\overline{X}_2 - \overline{X}_1$$

How variable is this estimate?

# Standard error for point estimate

Since our two samples are independent, so are our two sample means. The variance of their difference is the sum of their variances, $\sigma_1^2/n_1$ and $\sigma_2^2/n_2$.

Hence, the standard deviation of $\overline{X}_2 - \overline{X}_1$ is

$$\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

We estimate it with

$$SE_{\overline{X}_2 - \overline{X}_1} = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

# Example

The following are sample summary statistics for the birth weight data.

| smoke | mean | sd | n |
|---|---|---|---|
| nonsmoker | 7.18 | 1.43 | 100 |
| smoker | 6.78 | 1.60 | 50 |

Using these sample statistics, we estimate the difference in mean birth weight between smoking mothers and non-smoking mothers is $\overline{X}_{nonsmoker} - \overline{X}_{smoker} = 7.18 - 6.78 = 0.4$ pounds.

The standard error on this estimate is

$$SE_{\overline{X}_{nonsmoker} - \overline{X}_{smoker}} = \sqrt{\frac{s^2_{smoker}}{n_1} + \frac{s^2_{nonsmoker}}{n_2}} = \sqrt{\frac{1.6^2}{50} + \frac{1.43^2}{100}} = 0.27$$

# Sampling distribution for t-ratio

The quantity

$$t = \frac{(\overline{X}_2 - \overline{X}_1) - (\mu_2 - \mu_1)}{SE_{\overline{X}_2 - \overline{X}_1}}$$

is called Welch's two sample t-ratio.

Like in the one sample case, we have an approximately Normal point estimate, and are dividing by an estimated standard error.

It turns out the sampling distribution of this t-ratio is also close to a t-distribution but the degrees of freedom parameter is complicated, a.k.a Satterthwaite-Welch approximation to the degrees of freedom.

Sometimes people use the smaller of $n_1 - 1$ and $n_2 - 1$ as a quick to calculate, but conservative degrees of freedom. (The actual degrees of freedom will always be larger than this number, but no larger than $n_1 + n_2 - 2$)

We'll use R to find the degrees of freedom.

# Confidence intervals and tests

A $(1-\alpha)100\%$ confidence interval for the difference in population means, $\mu_2 - \mu_1$, is

$$(\overline{X}_2 - \overline{X}_1) \pm t_{\alpha/2}^{(df)} \times SE_{\overline{X}_2 - \overline{X}_1}$$

To test the null hypothesis that $\mu_2 - \mu_1 = \delta_0$, the statistic

$$\frac{(\overline{X}_2 - \overline{X}_1) - \delta_0}{SE_{\overline{X}_2 - \overline{X}_1}}$$

is compared to a t-distribution with Satterthwaite-Welch approximation degrees of freedom.

# Example of calculations

The null hypothesis the population mean weights are the same for smoking and non-smoking mothers is equivalent to:
$H_0 : \mu_{nonsmoker} - \mu_{smoker} = 0.$

The alternative is that the population mean weights are different for smoking and non-smoking mothers, $H_A : \mu_{nonsmoker} - \mu_{smoker} \neq 0$

Our test statistic is

$$t = \frac{(\overline{X}_2 - \overline{X}_1) - 0}{SE_{\overline{X}_2 - \overline{X}_1}} = \frac{0.4 - 0}{0.27} = 1.5$$

The degrees of freedom is somewhere between 49 and 148, which puts the critical value, $t_{0.975}$, somewhere between 2.01 and 1.97.

Without knowing the exact degrees of freedom we can still determine we would fail to reject the null hypothesis at the 5% significance level.

# Example of calculations

A 95% confidence interval for the difference in population means would be of the form

$$0.4 \pm t_{0.975}^{(df)} \times 0.27$$

with $t_{0.975}^{(df)}$ somewhere between 1.97 and 2.01. Using the larger value, 2.01, gives a slightly conservative interval of:

$$(-0.14, 0.94)$$

# Calculations in R

Notice the degrees of freedom is 89.277, so critical value is actually 1.99.

```
t.test(weight ~ smoke, data = births)
```

```
##
##  Welch Two Sample t-test
##
## data:  weight by smoke
## t = 1.4967, df = 89.277, p-value = 0.138
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.1311663  0.9321663
## sample estimates:
## mean in group nonsmoker    mean in group smoker
##                  7.1795                  6.7790
```

# Statistical Summary

*There is no evidence that the mean weight of babies born to smokers is different to the mean weight of babies born to nonsmokers (Welch's t-test, p-value = 0.14, df = 89.28).*

*With 95% confidence the mean weight of babies born to mothers who smoke is between 0.93 pounds lighter and 0.13 pounds heavier than the mean weight of babies born to nonsmokers*

Even though the difference in means was insignificant, it's important to still report a confidence interval.

Reporting the confidence interval allows a reader to determine if the plausible range includes values that may be of practical significance.

In this case, our plausible interval includes the possibility that babies born to smokers could be on average half a pound lighter than those born to non-smokers, a practically significant amount.

# Caveats

**Don't use this for paired data!**

A key assumption of the two sample t procedures is that the samples from the two populations are independent.

When there is pairing, this assumption is violated, and the procedure is inappropriate. You'll demonstrate this in lab/homework.

**There is another two sample t-test**

There is an alternative *equal variance* two sample t-test that makes the additional assumption the two populations have the same variance. This assumption is hard to verify and generally not true, so we don't recommend this version.