Agenda

1. Type I and type II error
2. Difference of paired samples
3. Difference of two means

Warm-up: One-sample Mean Problem

1. Should you generate electricity with your own personal wind turbine? That depends on whether you have enough wind on your site. To produce enough energy, your site should have an annual average wind speed above 8 mph. One candidate site was monitored for a year, with speeds recorded every 6 hours. A total of 1114 readings of wind speed averaged 8.019 mph with a standard deviation of 3.813 mph. What would you tell the landowner about whether this site is suitable for a small wind turbine? Test an appropriate hypothesis and construct a 99% confidence interval for your estimate.

Gifted Children's Parents Since in this data set, the IQ of both parents is recorded for all children, the IQ data is naturally paired.

```
library(openintro)
library(mosaic)
favstats(~motheriq, data = gifted)

##  min     Q1 median     Q3 max     mean       sd  n missing
##  101 113.75    118 122.25 131 118.1667 6.504943 36       0

favstats(~fatheriq, data = gifted)

##  min  Q1 median     Q3 max     mean      sd  n missing
##  110 112    115 116.25 126 114.7778 3.48147 36       0
```

We can define a new variable, `diff`, to be the difference between the mother's IQ and the father's for each gifted child.

```
gifted <- gifted %>%
  mutate(diff = motheriq - fatheriq)
favstats(~diff, data = gifted)

##  min   Q1 median   Q3 max     mean       sd  n missing
##  -15 -2.5    4.5 8.25  18 3.388889 7.453773 36       0
```

These are the two conditions for using a t-based sampling distribution for a single mean:

1. The samples come from independent observations
2. The distribution of the original variable is approximately normal, or the sample size is large

After checking our assumptions, we can answer the following questions:

1. Find a 90% confidence interval for the mean IQ of the mothers. Do the same for the fathers. Do they overlap?

2. Test the hypothesis that the mothers of gifted children have higher IQs, on average, than the fathers. Write out all of the steps. What do you conclude?

   (a) State the null and alternative hypotheses

   (b) Check that `diff` meets the conditions listed above

   (c) Compute the standard error of the mean ($SE_{\texttt{diff}}$) and the degrees of freedom

   (d) Compute the test statistic (t)

   (e) Compute the p-value and draw a conclusion [Use the `pt()` function in R.]

   (f) Write a sentence that provides an interpretation of your result

Practice Problem

1. A sample of n = 9 college students is used to evaluate the effectiveness of a new Study Skills Workshop. Each students grade point average (GPA) is recorded for the semester before the workshop and for the semester after the workshop. The average GPA improved by an average of 0.60 points, 95% CI: [.37, .83]. Based on the condence interval, is the Study Skills Workshop effective in improving students GPAs? Explain your reasoning.

2. Bonus: Find the SE of the mean difference based on the information provided.

Difference of two means Often the data are *not* naturally paired. In particular, we are often interested in comparing means from two groups of unequal sizes. For example, the 11 children whose fathers had higher IQs than mothers had a lower average score on the skills test than the 25 children whose mothers had higher IQs than the fathers.

```
favstats(score ~ (diff > 0), data = gifted)

##   (diff > 0) min    Q1 median  Q3 max     mean       sd  n missing
## 1      FALSE 150 152.5    156 161 164 156.5455 4.906397 11       0
## 2       TRUE 154 159.0    160 163 169 160.2800 4.097967 25       0
```

Now the samples are *not* naturally paired. How do we know if the observed difference in means between these two groups is meaningful? Let X be the random variable that gives the analytical skills test score for a gifted child whose father has a higher IQ than her mother, and let Y be the random variable that gives the test score for a gifted child whose mother has a higher IQ. Then we need to understand the sampling distribution of the test statistic $D = \bar{X} - \bar{Y}$.

Just as we did with proportions, the standard error of the difference is a combination of the standard errors of the variables.

$$SE_D = \sqrt{(SE_X)^2 + (SE_Y)^2}$$

If both X and Y meet the conditions for a t-based sampling distribution, then D will meet those conditions as well. We typically use $\min(n_1 - 1, n_2 - 1)$ for the degrees of freedom.

The hypothesis test for a difference of two means constructed in this manner is called the *two-sample t-test*, and it is a commonly applied statistical technique.

1. Use the information above to conduct a two-sample t-test for a difference in mean test score between gifted children whose fathers have higher IQs vs. those whose mothers have higher IQs.