Agenda

1. Inference for a single numerical mean

Warmup: Gifted Children An investigator is interested in understanding the relationship, if any, between the analytical skills of young gifted children and the following variables: father's IQ, mother's IQ, average number of hours per week the child watched an educational program on TV during the past three months, average number of hours per week the child watched cartoons on TV during the past three months. The analytical skills are evaluated using a standard testing procedure. Data were collected from schools in a large city on a set of 36 children who were identified as gifted children soon after they reached the age of four.

For 25 of the 36 children, the child's mother's IQ was higher than that of the father. Find a 95% confidence interval for the true proportion of gifted children whose mothers have higher IQs than their fathers.

Inference for a Mean We know how to make inferences about the value of a population proportion p, for a binary variable. The critical step was to construct an approximation of the sampling distribution of the sample proportion, $\hat{p}$. What if the the variable that we want to make inferences about is *numerical*? In this case we need to approximate the sampling distribution of the sample mean, $\bar{x}$. How can we do this?

Gifted Children's scores Use the information presented below to construct a 95% confidence interval for the mean analytical score among gifted children.

```
require(openintro)
require(mosaic)
favstats(~score, data = gifted)

##  min  Q1 median  Q3 max     mean       sd  n missing
##  150 155    159 162 169 159.1389 4.630043 36       0
```

1. Compute the standard error of the mean.

2. Find the appropriate critical value in the appropriate t-distribution. [Use the `qt` function in R.]

3. Write down the confidence interval.

4. Assume that the standard deviation presented above was actually the standard deviation of the scores in the whole population. Compute the confidence interval again, and compare the new interval to the one you found previously (using the t-distribution). Are they importantly different?

Inference for Paired Data: 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.

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?

Solution to Warmup We do the standard inference for a proportion:

```
require(openintro)
require(mosaic)
p_hat <- tally(~(fatheriq < motheriq), data = gifted, format = "proportion")[1]
se <- sqrt(p_hat * (1 - p_hat) / nrow(gifted))
p_hat + qnorm(c(0.025, 0.975)) * se

## [1] 0.5439707 0.8449182
```

Sampling Distribution of the Mean Let X be any random variable, with (population) mean $\mu_X = \mathbb{E}[X]$ and standard deviation $\sigma_X = sd(X)$.

- Suppose we take n observations of X. Then define a new random variable $\bar{X}$ which gives us the sample mean.
- The expectation of $\bar{X}$ is:

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n}\sum_{i=1} X_i\right] = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}[X_i] = \frac{1}{n}\cdot n \cdot \mathbb{E}[X] = \mu_X$$

- On the other hand, the variance of the sample mean decreases by a factor of n:

$$Var[\bar{X}] = Var\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \left(\frac{1}{n}\right)^2 \sum_{i=1}^{n} Var[X_i] = \frac{1}{n^2}\cdot n \cdot Var[X] = \frac{\sigma_X^2}{n}$$

- $SE_{\bar{X}} = \sqrt{Var[\bar{X}]} = \sigma_X/\sqrt{n}$
- Accordingly the standard deviation of the sample mean decreases by a factor of $\sqrt{n}$.
- Big Idea: If you want to halve the margin of error associated with your sample mean, you need to quadruple your sample size! ($\sigma_{\bar{x}} = \sigma_X/\sqrt{n}$).
- Recall the Central Limit Theorem: In many situations, if n is large (at least 30 or 40), then the sample mean is approximately normally distributed **regardless** of the underlying distribution of X.
- Thus, if we know the population standard deviation and either the sample size is very large, or the population is normally distributed, the sampling distribution of the mean is approximately normal. Consider

$$z = \frac{\bar{x} - \mu_X}{\sigma_X/\sqrt{n}} = \frac{\text{sample mean} - \text{population mean}}{SE},$$

  this test statistic follows a standard normal distribution. Thus, we construct confidence intervals for the population mean of the form:

$$\bar{x} \pm z^*_{\alpha/2} \cdot \frac{\sigma_X}{\sqrt{n}}.$$

- In reality, we rarely know σ_X, so we approximate it with s_X, the standard deviation of the sample. The test statistic thus becomes

$$t = \frac{\bar{x} - \mu_X}{s_X/\sqrt{n}} = \frac{\text{sample mean} - \text{population mean}}{SE},$$

You can prove mathematically that if X is normally distributed, then this test statistic follows a t-distribution on $n-1$ degrees of freedom. This leads to confidence intervals of the form

$$\bar{x} \pm t^*_{\alpha/2} \cdot \frac{s_X}{\sqrt{n}},$$

where $t^*_{\alpha/2}$ is a value from the t-distribution with $n-1$ degrees of freedom (accessed in R by `qt()`).

- The t-distribution is similar to the Standard Normal distribution (unimodal, symmetric), but is indexed by an additional parameter (the degrees of freedom), and has fatter tails. In fact, the t-distribution converges to the Standard Normal for an infinite sample size (i.e. $t(\infty) = N(0,1)$).
- This procedure still assumes that X is normally distributed, and forces the confidence intervals to be symmetric.

```
plotDist("norm", lwd=5)
plotDist("t", params = list(df=1), add=TRUE, col="red")
plotDist("t", params = list(df=2), add=TRUE, col="red")
plotDist("t", params = list(df=4), add=TRUE, col="red")
plotDist("t", params = list(df=8), add=TRUE, col="red")
plotDist("t", params = list(df=16), add=TRUE, col="red")
plotDist("t", params = list(df=32), add=TRUE, col="red")
plotDist("t", params = list(df=64), add=TRUE, col="red")
plotDist("t", params = list(df=128), add=TRUE, col="red")
plotDist("t", params = list(df=1024), add=TRUE, col="red")
```

Solution to Gifted Children's scores Here we compare the t-based interval to the z-interval:

```
x_bar <- mean(~score, data = gifted)
n <- nrow(gifted)
se <- sd(~score, data = gifted) / sqrt(n)
x_bar + qnorm(c(0.025, 0.975)) * se

## [1] 157.6264 160.6513

x_bar + qt(c(0.025, 0.975), df = (n - 1)) * se

## [1] 157.5723 160.7055
```

Solution to Gifted Children's Parents We need to make three intervals:

```
# mothers
mean_m <- mean(~motheriq, data = gifted)
se_m <- sd(~motheriq, data = gifted) / sqrt(n)
mean_m + qt(c(0.05, 0.95), df = (n - 1)) * se_m

## [1] 116.3349 119.9984

# fathers
mean_m <- mean(~fatheriq, data = gifted)
se_m <- sd(~fatheriq, data = gifted) / sqrt(n)
mean_m + qt(c(0.05, 0.95), df = (n - 1)) * se_m
```

```
## [1] 113.7974 115.7581

# pairs
gifted <- transform(gifted, diff = motheriq - fatheriq)
x_bar <- mean(~diff, data = gifted)
se <- sd(~diff, data = gifted) / sqrt(n)
2 * pt(x_bar/se, df = (n-1), lower.tail = FALSE)

## [1] 0.009897971
```