Uncertainty, 1

Frank Edwards

The basic challenge: Will Obama win Florida in '08?
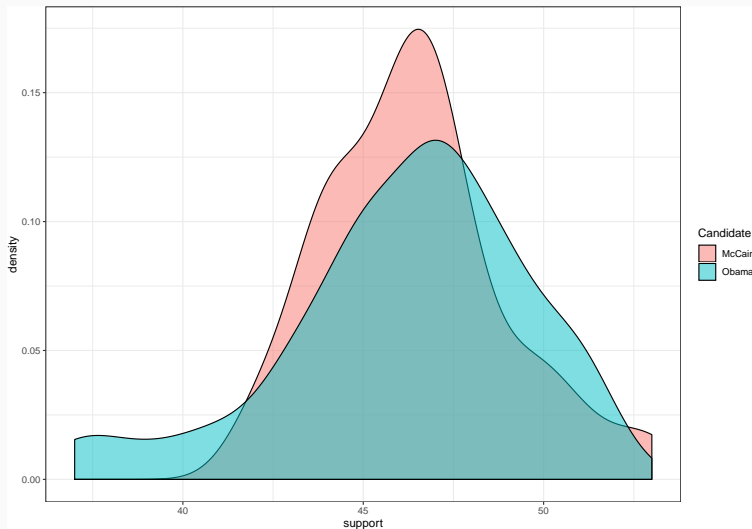
- Take a simple random sample of n voters
- Ask them their preferences (support for Obama $x = [0, 1]$)
- Compute the proportion of voters supporting Obama in the sample
  $$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$
- Use $\bar{x}$ to inform our understanding of p, the true proportion of voters supporting Obama

$$\text{estimation error} = \text{estimate} - \text{truth} = \bar{x} - p$$

$$\text{estimation error} = \text{estimate} - \text{truth} = \bar{x} - p$$

Problems?

The data:

$$x \sim \text{Bernoulli}(p)$$

```
sample <- rbinom(size = 1, n = 10, prob = 0.55)
### x_i
sample
```

```
## [1] 1 1 1 0 1 0 1 0 1 1
```

```
### \bar{x}
sum(sample)/length(sample)
```

```
## [1] 0.7
```

What is the estimation error?

Recall that a Binomial variable represents the number of successes from *n*
Bernoulli trials with a probability of success *p*

Recall that a Binomial variable represents the number of successes from *n* Bernoulli trials with a probability of success *p*

We can represent the average share of Obama supporters in a sample of voters as a Binomial random variable:

$$\bar{x} \times n \sim \text{Binomial}(p, n)$$

Recall that a Binomial variable represents the number of successes from *n* Bernoulli trials with a probability of success *p*

We can represent the average share of Obama supporters in a sample of voters as a Binomial random variable:

$$\bar{x} \times n \sim \text{Binomial}(p, n)$$

What does this probability distribution mean?

```
sample1 <- rbinom(size = 1, n = 10, prob = 0.55)
sample2 <- rbinom(size = 1, n = 10, prob = 0.55)
sample3 <- rbinom(size = 1, n = 10, prob = 0.55)
mean(sample1)
```

```
## [1] 0.5
```

```
mean(sample2)
```
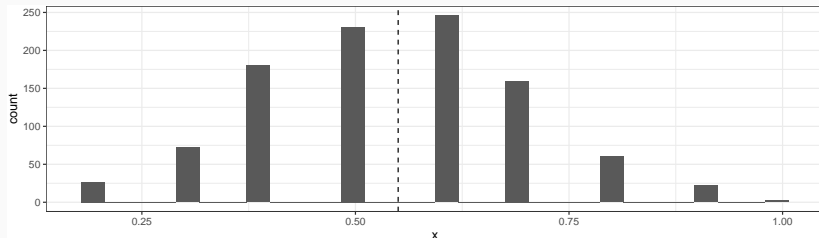
```
## [1] 0.4
```

```
mean(sample3)
```

```
## [1] 0.7
```

Recall the Law of Large Numbers: as a sample of draws from a random variable increases, the sample mean converges to the population mean $E(X)$
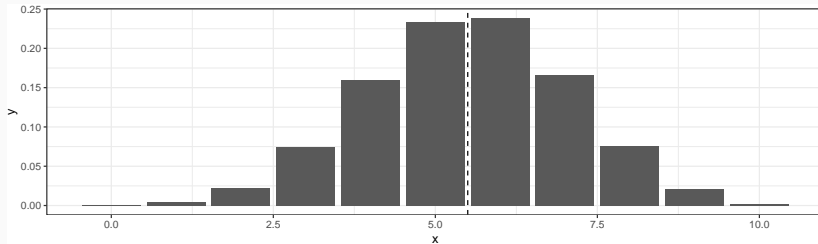
$$\bar{x}_n \rightarrow E(X)$$

```r
sim_dat <- data.frame(x = rep(NA, 1000))
for (i in 1:1000) {
    sample <- rbinom(size = 1, n = 10, prob = 0.55)
    sim_dat$x[i] <- mean(sample)
}
ggplot(sim_dat, aes(x = x)) + geom_histogram(bins = 30) + geom_vline(xintercept = 0.55,
    lty = 2)
```

Compare this to Binomial(n,p)

```r
x <- data.frame(x = 0:10, y = dbinom(0:10, size = 10, prob = 0.55))
ggplot(x, aes(x = x, y = y)) + geom_col() + geom_vline(xintercept = 0.55 * 10,
    lty = 2)
```

$$\bar{x}_n \rightarrow E(X)$$

- So if we resample *x* from the same population a large number of times, we approximate the *sampling distribution* of $\bar{x}$
- This holds under simple random sampling, where each unit has an equal probability of being selected

So is our sample biased?

$$\text{estimation error} = \text{estimate} - \text{truth} = \bar{x} - p$$

$$\text{bias} = E(\text{estimation error}) = E(\text{estimate} - \text{truth}) = E(\bar{x}) - p$$

Because $E(\bar{x}) = p$ under simple random sampling

$$\text{bias} = E(\bar{x}) - p = p - p = 0$$

The sample mean $\bar{x}$ is, on average, an *unbiased* estimator of the population average p as long as we've obtained a simple random sample.

So is our sample biased?

$$\text{estimation error} = \text{estimate} - \text{truth} = \bar{x} - p$$

$$\text{bias} = E(\text{estimation error}) = E(\text{estimate} - \text{truth}) = E(\bar{x}) - p$$

Because $E(\bar{x}) = p$ under simple random sampling

$$\text{bias} = E(\bar{x}) - p = p - p = 0$$

The sample mean $\bar{x}$ is, on average, an *unbiased* estimator of the population average p as long as we've obtained a simple random sample.

However: any individual sample has unobservable error $\bar{x_n} - p$

12

An estimator is biased when $E(\bar{x}) \neq p$

- Bias means that the expectation is systematically different from the true value
- Positive bias means that $E(\bar{x}) > p$, negative bias means $E(\bar{x}) < p$

An estimator is consistent if it converges to the parameter as n increases, or

$$\bar{x}_n \to E(X)$$

- Simple random sampling yields unbiased and consistent estimators of a parameter

- Simple random sampling yields unbiased and consistent estimators of a parameter
- In experimental settings: Random assignment yields an unbiased estimator of the Sample Average Treatment Effect
- In experimental settings: Simple random sampling combined with random assignment yields an unbiased and consistent estimator of the Population Average Treatment Effect
- See pages 317-322 for more discussion of RCTs and bias in estimators

How wrong are we, on average?

For our vote share example, pretend that $p = 0.55$, and $n = 10$. For a Binomial random variable expressed as a proportion:

$$\sigma^2 = \frac{p(1-p)}{n} = \frac{0.55(1-0.55)}{10} \approx 0.025$$

We define a standard error as the standard deviation of a sampling distribution

$$SE = \sigma = \sqrt{0.025} \approx 0.16$$

How wrong are we, on average?

For our vote share example, pretend that $p = 0.55$, and $n = 10$. For a Binomial random variable expressed as a proportion:

$$\sigma^2 = \frac{p(1-p)}{n} = \frac{0.55(1-0.55)}{10} \approx 0.025$$

We define a standard error as the standard deviation of a sampling distribution

$$SE = \sigma = \sqrt{0.025} \approx 0.16$$

When we sample 10 voters, our estimate of the share of the population that supports Obama will have an average error of about 0.16, or 16 percentage points!

How wrong are we, on average?

For our vote share example, pretend that $p = 0.55$, and $n = 10$. For a Binomial random variable expressed as a proportion:

$$\sigma^2 = \frac{p(1-p)}{n} = \frac{0.55(1 - 0.55)}{10} \approx 0.025$$

We define a standard error as the standard deviation of a sampling distribution

$$SE = \sigma = \sqrt{0.025} \approx 0.16$$

When we sample 10 voters, our estimate of the share of the population that supports Obama will have an average error of about 0.16, or 16 percentage points!

What happens as we increase n?

In general, the standard error of a sample mean $\bar{x}_n$ where x is a sample of n i.i.d random variables is:

$$SE(\bar{x}) = \frac{sd_x}{\sqrt{n}}$$

We use the observed data to approximate the standard deviation of the sampling distribution for a population parameter, such as a mean.

It tells us how far away $\bar{x}$ is from $E(x)$ on average

Describing a distribution

Let's return to our voter polls. Assume each of these polls is a simple random sample of voters.

```
data(polls08)
fl <- polls08 %>% filter(state == "FL") %>% select(Obama)

head(fl)
```

```
##   Obama
## 1    48
## 2    50
## 3    47
## 4    47
## 5    51
## 6    50
```

Each of these represents a draw from the sampling distribution of voters in Florida. The vote share estimate is the mean of a group of sampled voters, whose preference for Obama is a Bernoulli variable with probability *p*

The sum of i.i.d. Bernoulli trials is Binomial(n,p). When we add i.i.d. Binomial random variables together, the Central Limit Theorem shows that they converge toward a Normal distribution as the number of variables increases.

```
ggplot(fl, aes(x = Obama)) + geom_histogram()
```
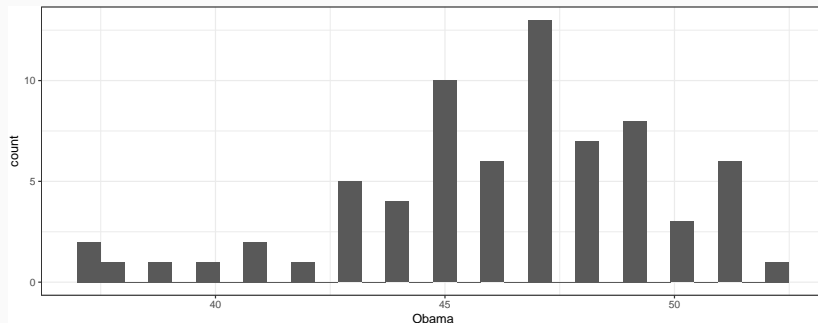
More formally:

$$\bar{x}_n \sim N\left(E(x), \frac{V(x)}{n}\right) \text{ as } n \to \infty$$

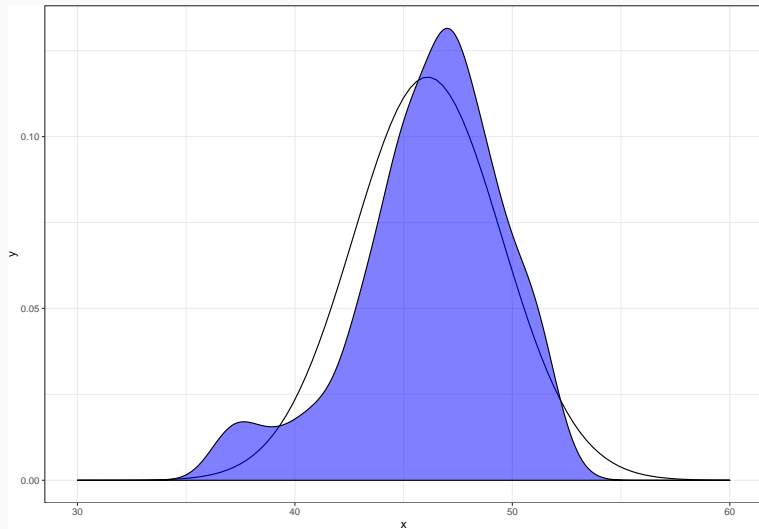The sampling distribution of $\bar{x}$ is approximately Normal, if we resample x many times.

More formally:

$$\bar{x}_n \sim N\left(E(x), \frac{V(x)}{n}\right) \text{ as } n \to \infty$$

The sampling distribution of $\bar{x}$ is approximately Normal, if we resample x many times.

Using our polling data, we obtained a distribution of $\bar{x}$ with a mean of 46.1 and standard error of 3.4

Observed sample means (blue), and N(46.1, 3.4)

To specify where we think a new sample would land, we can construct a confidence interval.

1. Choose a a confidence level (e.g. 0.9, 0.95, 0.99)
2. Compute the quantile of the Normal distribution that corresponds to that confidence level (symmetric)
3. Multiply the value of this quantile by the standard error of the quantity, then add and subtract this value from the sample mean

Confidence interval example

Obtain a 90 percent confidence interval of Obama vote share in Florida (2008)

$$\alpha = 1 - 0.9 = 0.1$$

Obtain a 90 percent confidence interval of Obama vote share in Florida (2008)

$$\alpha = 1 - 0.9 = 0.1$$

Because a Normal is symmetric, we want to obtain all of the probability mass of the Normal curve above the 0.05 quantile and below the 0.95 quantile, or $0 + \alpha/2$ and $1 - \alpha/2$.

On the *z* score scale, this corresponds to

```
qnorm(0.05, 0, 1)
```

```
## [1] -1.644854
```

```
qnorm(0.95, 0, 1)
```

```
## [1] 1.644854
```

Our 90% confidence interval is equal to:

$$\bar{x} \pm z_{\alpha/2} \times SE$$

There is a 90 percent chance that if we draw a new sample of voters, it's mean Obama vote share would fall in this interval

```
mean(fl$Obama) - 1.64 * sd(fl$Obama)
```

```
## [1] 40.66452
```

```
## and
mean(fl$Obama) + 1.64 * sd(fl$Obama)
```

```
## [1] 51.70168
```

Now we need the z-score for a Normal CDF at 0.975 and 0.025, and obtain our interval as before

```
## 90 percent interval
probs <- qnorm(c(0.05, 0.95), 0, 1)
mean(fl$Obama) + probs * sd(fl$Obama)
```
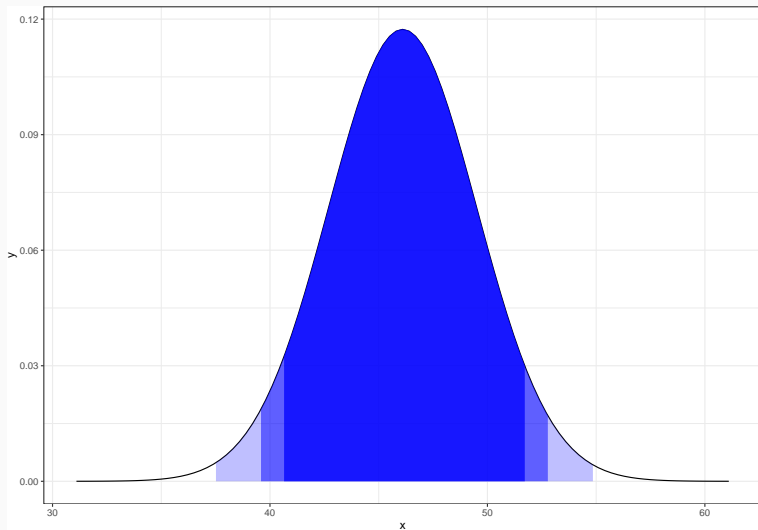
```
## [1] 40.64819 51.71801
```

```
## 95 percent interval
probs <- qnorm(c(0.025, 0.975), 0, 1)
mean(fl$Obama) + probs * sd(fl$Obama)
```

```
## [1] 39.58784 52.77835
```

What do you notice?

Confidence intervals for $\alpha = 0.1, 0.05, 0.01$

What does this interval mean?

```
## 95 percent interval
probs <- qnorm(c(0.025, 0.975), 0, 1)
mean(fl$Obama) + probs * sd(fl$Obama)
```

```
## [1] 39.58784 52.77835
```

What does this interval mean?

```
## 95 percent interval
probs <- qnorm(c(0.025, 0.975), 0, 1)
mean(fl$Obama) + probs * sd(fl$Obama)


## [1] 39.58784 52.77835
```

It means that if we draw another sample from this same population, the sampled value will fall within the interval 95 percent of the time.

What does this interval mean?

```
## 95 percent interval
probs <- qnorm(c(0.025, 0.975), 0, 1)
mean(fl$Obama) + probs * sd(fl$Obama)
```

```
## [1] 39.58784 52.77835
```

It means that if we draw another sample from this same population, the sampled value will fall within the interval 95 percent of the time.

It does *not* mean that probability that the true value of the population mean falls within the interval is 0.95.

What does this interval mean?

```
## 95 percent interval
probs <- qnorm(c(0.025, 0.975), 0, 1)
mean(fl$Obama) + probs * sd(fl$Obama)
```

```
## [1] 39.58784 52.77835
```

It means that if we draw another sample from this same population, the sampled value will fall within the interval 95 percent of the time.
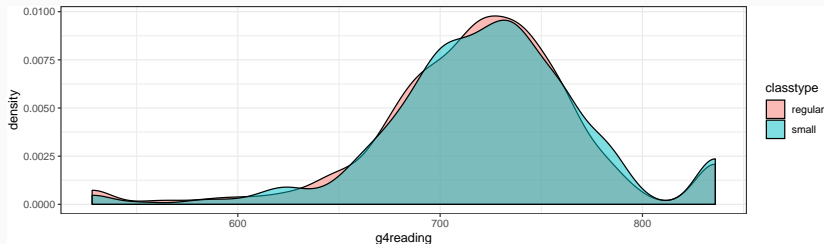
It does *not* mean that probability that the true value of the population mean falls within the interval is 0.95.

It does mean that when we draw a sample and construct a 95 percent interval, it will contain the population mean 95 percent of the time.

Let's return to the class-size experiment. We can use a confidence interval to provide more information about the location of

```
data(STAR)
STAR <- STAR %>% mutate(classtype = ifelse(classtype == 1, "small", ifelse(classtype ==
    2, "regular", NA))) %>% filter(!(is.na(classtype)), !(is.na(g4reading)))
ggplot(STAR, aes(x = g4reading, fill = classtype)) + geom_density(alpha = 0.5)
```

```r
STAR_groups <- STAR %>% group_by(classtype) %>% summarise(xbar_g4reading = mean(g4reading),
    se_g4reading = sd(g4reading)/sqrt(n()))
```

And 95% Confidence Intervals for each group

```r
alpha = 0.05
STAR_groups %>% mutate(ci_lower = xbar_g4reading - qnorm(1 - alpha/2, 0, 1) *
    se_g4reading, ci_upper = xbar_g4reading + qnorm(1 - alpha/2, 0, 1) * se_g4reading)
```

```
## # A tibble: 2 x 5
##   classtype xbar_g4reading se_g4reading ci_lower ci_upper
##   <chr>              <dbl>        <dbl>    <dbl>    <dbl>
## 1 regular             720.         1.84     716.     723.
## 2 small               723.         1.91     720.     727.
```

```
SATE <- STAR_groups$xbar_g4reading[2] - STAR_groups$xbar_g4reading[1]
SATE_se <- sqrt(STAR_groups$se_g4reading[1]^2 + STAR_groups$se_g4reading[2]^2)
### 95 percent CI
SATE + qnorm(1 - alpha/2, 0, 1) * SATE_se
```
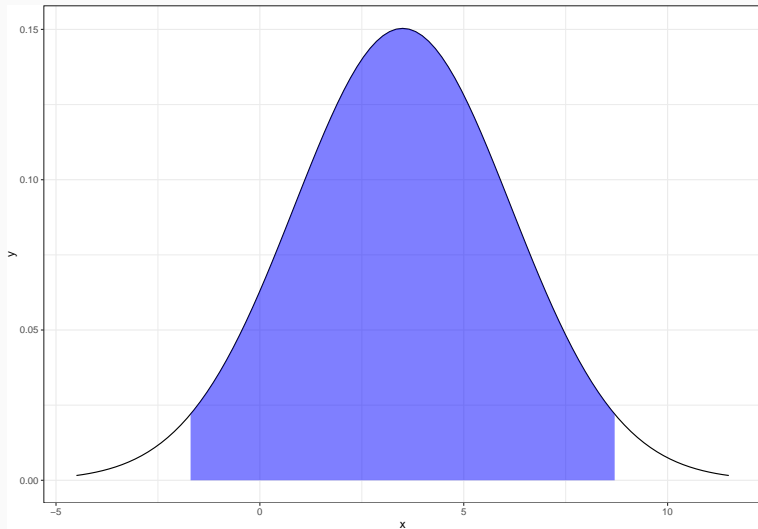
```
## [1] 8.701968
```

```
SATE - qnorm(1 - alpha/2, 0, 1) * SATE_se
```

```
## [1] -1.699503
```

We have 90 percent certainty that the true SATE falls within the interval $[-1.7, 8.7]$.

The estimated sampling distribution of the SATE, with 95 percent interval

We would like to know whether assignment to a small class in Kindergarten improves reading scores in fourth grade.

We start by proposing two hypotheses:

- H_0: Kindergarten class assignment has no effect on fourth grade reading scores
- H_1: Kindergarten class assignment has an effect on fourth grade reading scores

We will use the sampling distribution of the SATE to test this hypothesis

1. Specify a null and alternative hypothesis
2. Choose a test statistic and level of α
3. Derive the sampling distribution of the test statistic under the null hypothesis
4. Compute a p-value (the probability that, under the null hypothesis, we would observe a result as extreme as we observed in the data)
5. Reject the null hypothesis if $p \leq \alpha$, otherwise retain the null hypothesis

$$H_0 : \overline{x}_{small} - \overline{x}_{regular} = 0$$

$$H_1 : \overline{x}_{small} - \overline{x}_{regular} \neq 0$$

- Under the null hypothesis, we assume that the SATE has a Normal distribution with mean 0 and standard deviation equal to the observed standard error of the SATE
- Let's go with the conventional $\alpha = 0.05$ threshold (there's nothing special about it though!)
- Then we estimate the probability that we could observe the data under the null hypothesis using the Normal Cumulative Distribution Function (pnorm in R)

We will reject the null hypothesis if our test statistic falls in either of the red areas (two-tailed test):
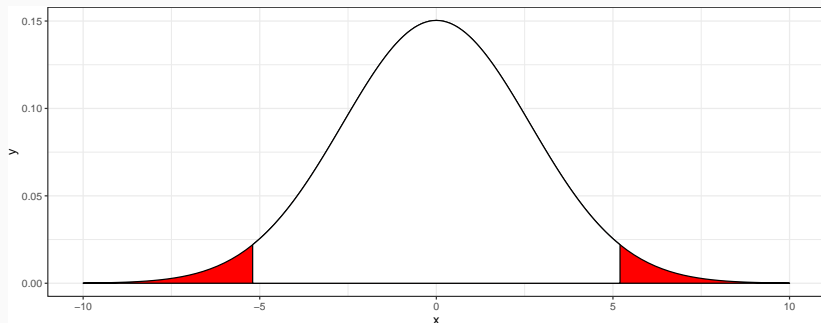
Because the Normal CDF is symmetric, we use the left-tail to estimate the probability that it falls within the rejection zone on either side

```
2 * pnorm(-abs(SATE), 0, SATE_se)
```

```
## [1] 0.1870072
```

Can we reject H_0?

R has built in functions for *t* tests. Note that, when we assume our variable is Normally distributed, we may use a *t* distribution rather than a Normal distribution for hypothesis tests

```
STAR_small <- STAR %>% filter(classtype == "small")
STAR_regular <- STAR %>% filter(classtype == "regular")
t.test(STAR_small$g4reading, STAR_regular$g4reading)
```

```
##
##  Welch Two Sample t-test
##
## data:  STAR_small$g4reading and STAR_regular$g4reading
## t = 1.3195, df = 1541.2, p-value = 0.1872
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -1.703591  8.706055
## sample estimates:
## mean of x mean of y
##  723.3912  719.8900
```

Words of warning

- Hypothesis tests and confidence intervals are tricky to interpret, and are often intrepreted incorrectly
- Hypothesis testing has become a *very* problematic feature of modern social science
- Just because a result is has a p-value of less than 0.05 does not mean that a finding is true!
- The routines of modern science exacerbate these problems

- Hypothesis tests and confidence intervals are tricky to interpret, and are often intrepreted incorrectly
- Hypothesis testing has become a *very* problematic feature of modern social science
- Just because a result is has a p-value of less than 0.05 does not mean that a finding is true!
- The routines of modern science exacerbate these problems

If we test 10 hypotheses in an analysis, the likelihood that at least one will be significant is:

```
1 - 0.95^10
```

```
## [1] 0.4012631
```

- We'll re-examine the Pager criminal record experiment armed with our new methods to evaluate uncertainty in sample statistics.
- The homework and data are in ~/hw/criminal_record/
- In lab, we'll get started on the homework
- Next week is uncertainty in linear regression