Significance tests

S520

October 13, 2016

Reference: Trosset chapter 9

Confidence intervals

Load the choral singer data:

```
singer = read.table("singer.txt", header=TRUE)
```

Let's find a 90% CI for the mean.

```
xbar = mean(singer$height)
q = qnorm(.95)
s = sd(singer$height)
n = length(singer$height)
# Lower bound
xbar - q * s/sqrt(n)
```

```
## [1] 66.88748
```

```
# Upper bound
xbar + q * s/sqrt(n)
```

```
## [1] 67.70826
```

Find a 80% CI for the proportion 6 feet or taller.

```
phat = sum(singer$height >= 72) / n
q = qnorm(.90)
# Lower bound
phat - q * sqrt(phat*(1 - phat)/n)
```

```
## [1] 0.1269982
```

```
# Upper bound
phat + q * sqrt(phat*(1 - phat)/n)
```

```
## [1] 0.1878954
```

Do confidence intervals work?

Let's do a simulation.

```
p = 0.2
phats = rbinom(10000,235,p)/235
lower.bounds = phats - q * sqrt(phats*(1 - phats)/n)
upper.bounds = phats + q * sqrt(phats*(1 - phats)/n)
summary(lower.bounds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.05436 0.15070 0.16660 0.16650 0.18260 0.26370
```

```
summary(upper.bounds)
```

```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
## 0.09883 0.21530 0.23340 0.23310 0.25150 0.34050
```

```
sum(p < lower.bounds) # should be 10% of the time
```

```
## [1] 825
```

```
sum(p > upper.bounds) # should be 10% of the time
```

```
## [1] 1063
```

It's a bit off, but close.

Significance testing

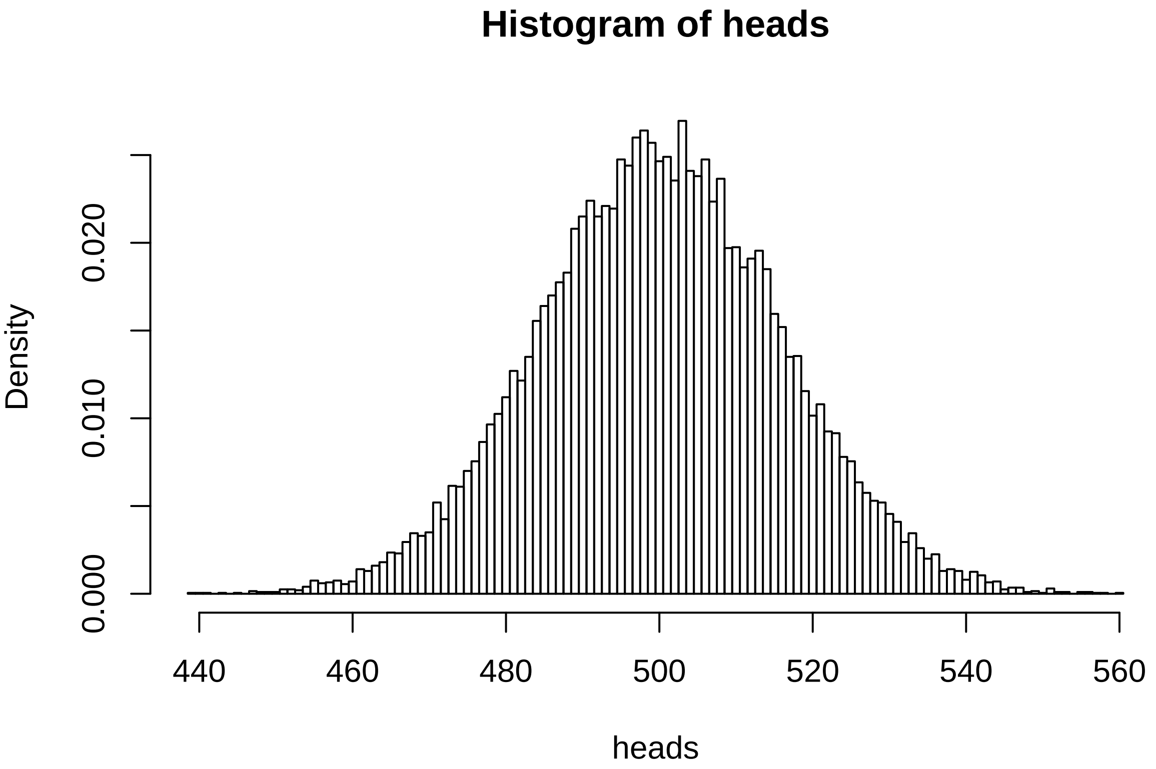
I toss a coin 1000 times and get 489 heads Is this enough to show that it's biased?

We can do a simulation:

```
heads = rbinom(20000, 1000, 0.5)
summary(heads)
```

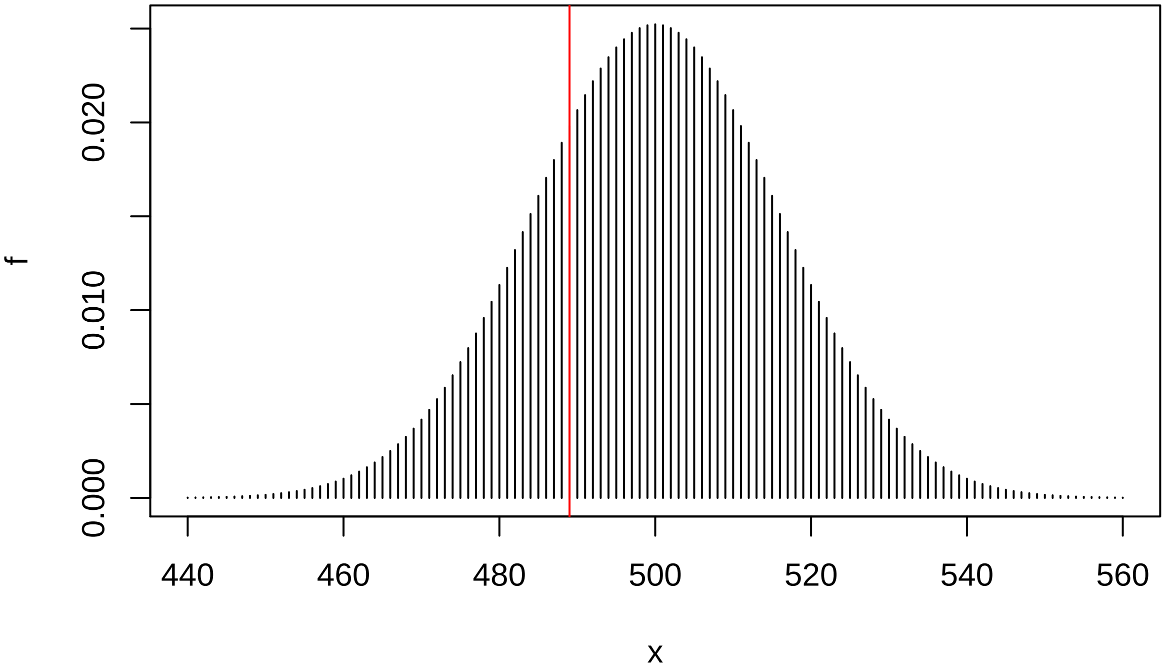
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   439.0   490.0   500.0   500.1   511.0   560.0
```

```
hist(heads, prob=T,
     breaks=(min(heads)-0.5):(max(heads)+0.5))
```



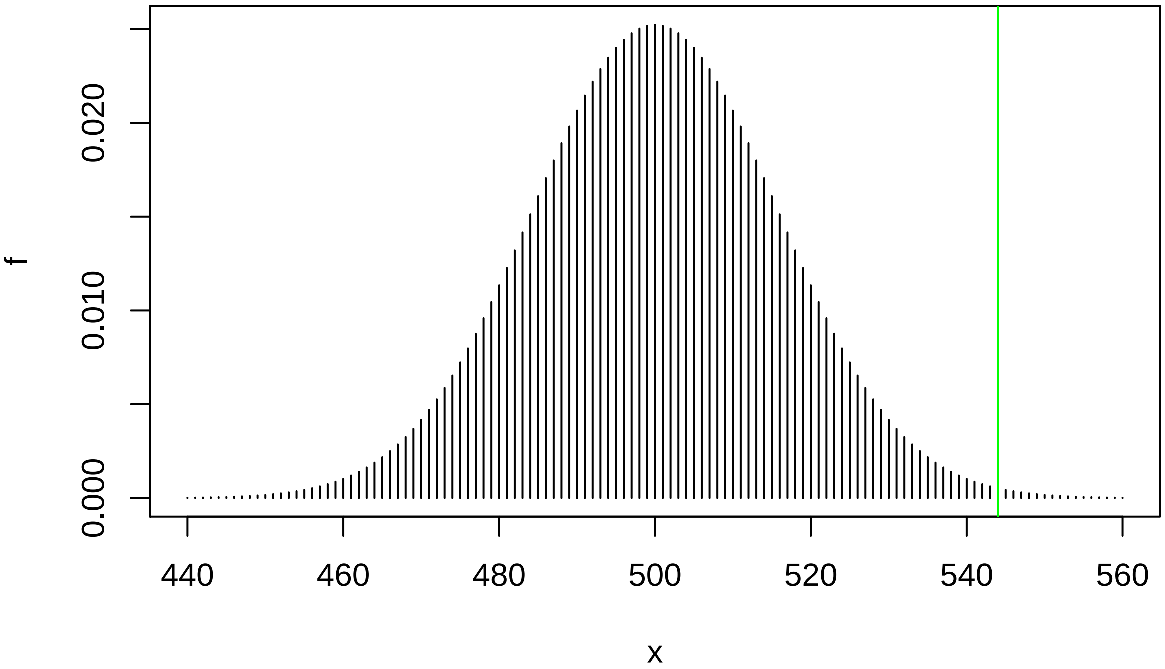
Or we could do exact calculations using the binomial distribution:

```
x = 440:560
f = dbinom(x, 1000, 0.5)
plot(x, f, type="h")
abline(v = 489, col="red")
```



It seems 489 heads is not at all unusual. But what about 544?

```r
plot(x, f, type="h")
abline(v = 544, col="green")
```



This is unusual. How unusual? We could find the probability of 544 or more heads:

```r
1 - pbinom(543, 1000, 0.5)
```

```
## [1] 0.002955377
```

But if 44 or heads more than expected counts as “unusual”, then 44 or more heads less than expected should also count as unusual. So it’s better to find the probability of being at least 44 heads from 500:

```r
pbinom(456, 1000, 0.5) + 1 - pbinom(543, 1000, 0.5)
```

```
## [1] 0.005910755
```

This is a small probability. So “544 heads in 1000 tosses” is not very consistent with the hypothesis of a fair coin.

Significance testing: The steps

1. Write down the null hypothesis (H_0) and the alternative hypothesis (H_1). Choose a **test statistic** that you can find the distribution for when the null hypothesis is true (plus any other necessary assumptions.)
2. Write down the **sampling distribution** for the test statistic, assuming the null hypothesis model is true. Decide what tails of the sampling distribution will count as “extreme.”
3. Collect random data. Use the observed data to calculate the test statistic.
4. Find the **significance probability** (P-value.)
5. Give a conclusion: Does the data fit the null hypothesis? The smaller the P-value, the less compatible the data is with the null.
6. It’s usually useful to find a (two-sided) confidence interval as well.

Optional steps (not personally recommended by the lecturer)

If you need to make a binary decision between your null and alternative hypotheses:

1a. Before collecting the data, ask yourself:

"Suppose the null hypothesis is true. What is the maximum probability of wronging choosing the alternative I am willing to accept?"

This is your **significance level** (denoted α).

5a. If the P-value turns out to be less than α, reject the null hypothesis in favor of the alternative. If the P-value turns out to be greater than α, do not reject the null hypothesis.

Some of the ideas above are confusing! Let's clarify a few issues.

What's the null? What's the alternative?

The null hypothesis is what you initially assume to be true, for the purpose of doing probability calculations. In chapter 10, we're doing tests about a population mean μ, so the null hypothesis will be a statement about μ.

A **two-tailed test** has hypotheses of the form:

$$H_0 : \mu = 10$$
$$H_1 : \mu \neq 10$$

For these hypotheses, both $\bar{x}$ much higher and $\bar{x}$ much lower than 10 will be incompatible with the null hypothesis (and compatible with the alternative.)

There are two kinds of **one-tailed test**. Here's a **right-tailed test:**

$$H_0 : \mu \leq 25$$
$$H_1 : \mu > 25$$

In the above example, only values of $\bar{x}$ well above 25 will be incompatible with the null.

Now for a **left-tailed test:**

$$H_0 : \mu \geq 60$$
$$H_1 : \mu < 60$$

Here, only values of $\bar{x}$ well below will be incompatible with the null.

What's a P-value?

The P-value is the probability under the null hypothesis model that the test statistic is as extreme or more extreme than the one observed.

When calculating the P-value, we always start from the assumption that the null is true. For a right-tailed test, the P-value is the probability under the null of a test statistic greater than or equal to the one observed. For a left-tailed test, the P-value is the probability under the null of a test statistic greater than or equal to the one observed. For a two-tailed test with a symmetric sampling distribution, take the smaller of the two probabilities above and double it.

Wait, what about comparing the P-value to 0.05?

The most common choice of significance level is $\alpha = 0.05$. This can be a good benchmark for doing simulations. However, it's a terrible idea to blindly use $\alpha = 0.05$ for every test. Here are some reasons you should avoid fixed significance levels (and to complain about fixed significance levels when they're forced on you):

- If you must make a decision, there's no reason to set the same threshold for every decision – and different people might want to use different thresholds. Always state your P-value (don't just say "significant" or "not significant.")
- If the P-value is bigger than 0.05, it's very tempting to say that the null hypothesis is true (or that there's no effect.) But this isn't the case – all you can say is the data is compatible with the null hypothesis. But the data may be compatible with many other hypotheses as well! You need more information to decide whether the null is (approximately) true or not.
- If the P-value is smaller than 0.05, it's very tempting to say that you've discovering something big and important. But the P-value doesn't tell you how big the effect is (that's what a confidence interval is for.) A P-value for a test of $\mu = 0$ tells you how compatible the data is with the hypothesis that the population mean is zero. If μ isn't zero, the test doesn't tell you what the mean *is*: regardless of the P-value, the mean could be 0.00001 or it could be a million.
- If the P-value is smaller than 0.05, it's tempting to say something like "there's a 95% chance the difference is real." This is wrong. Either the effect is real or it isn't – the hypothesis itself isn't random, so there's no frequentist probability that the null hypothesis is true. To make a probability statement about a hypothesis, you need to use subjective probability, well, it's subjective, so you're on your own.

Example: Do beautiful parents have more daughters?

In the general population: 48.5% of births are girls. However, an evolutionary psychologist has suggested that beautiful people are more likely to give birth to girls. Among People's Most Beautiful People, there were 157 girls out of 329 children. Is there evidence that People's Most Beautiful People give birth to girls at a higher rate than the rest of the population?

Hypotheses:

$$H_0 : p \leq 0.485$$
$$H_1 : p > 0.485$$

As our test statistic, we could use either the number of girls in the sample or (equivalently) the sample proportion of girls. We'll pick the number because that has an easy-to-deal-with binomial distribution, with $n = 329$ and $p = 0.485$.

The observed value of the test statistic is 157 girls. The P-value will be the "greater than or equal to" probability. We can calculate this using `pbinom()`:

```r
1 - pbinom(156, 329, 0.485)
```

```
## [1] 0.6321467
```

This is *not* a small P-value, so we have no evidence against the null hypothesis: that is, there's no real evidence that Most Beautiful People give birth to a higher proportion of girls than the rest of the population. (The Central Limit Theorem would give a similar result.)

If we were wanted to make a decision, we should have specified an α-level in advance. However, since we never choose an α-level bigger than 0.1 or so, we can safely say we do not reject the null hypothesis.

It's a good idea to give a 95% confidence interval as well:

```
phat = 157/329
se = sqrt(phat * (1-phat) / 329)
# Lower bound
phat - qnorm(.975) * se
```

```
## [1] 0.4232317
```

```
phat + qnorm(.975) * se
```

```
## [1] 0.5311756
```

The interval runs from 42% to 53%, which is pretty wide. Exercise: How many births would you need to get the width of the interval down to 5%?