# Chapter 9, part 1: Significance testing

*S520 online*

These notes are written to accompany Trosset chapter 9.1, 9.3, and 9.4.

### Example: Is this coin biased?

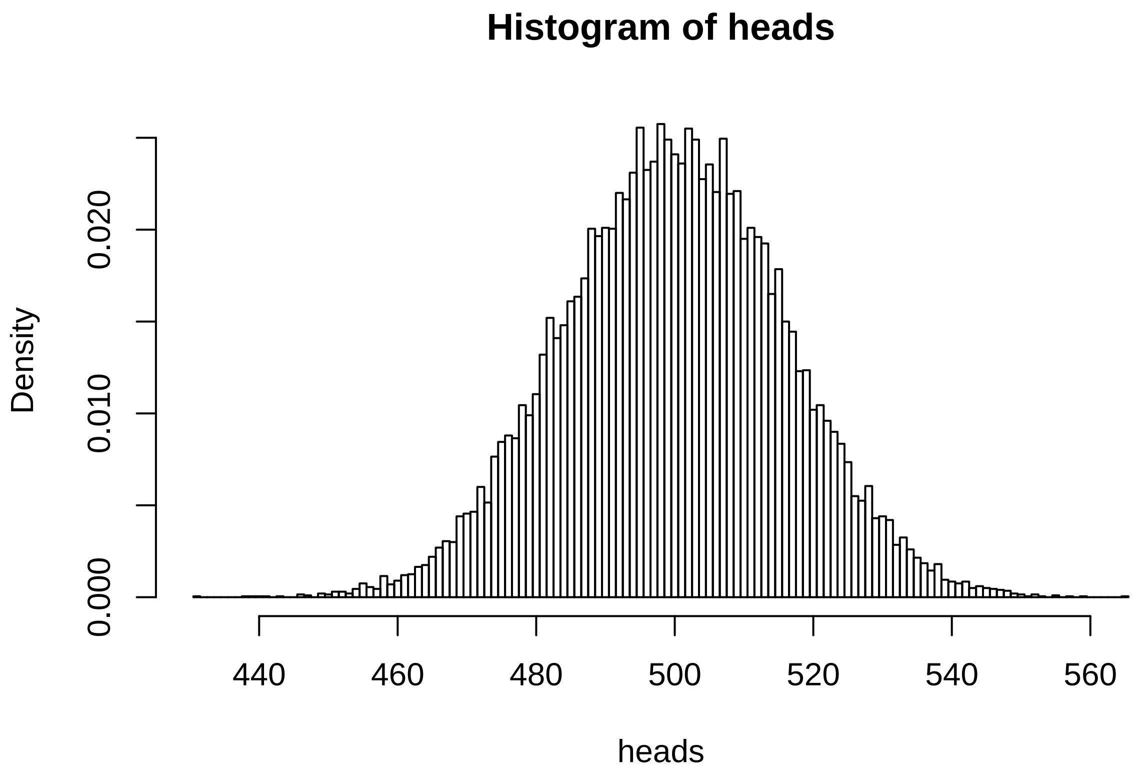
I toss a coin 1000 times and get 489 heads. Is this enough to show that it's biased?

We can do a simulation:

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

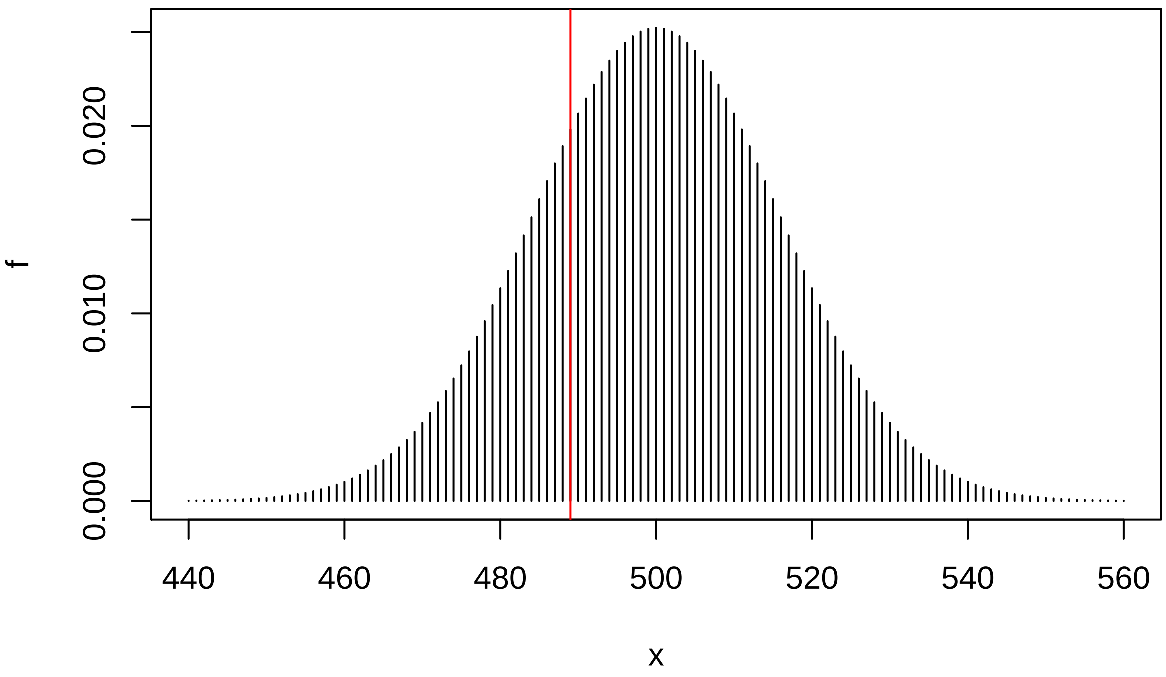
```
##    Min. 1st Qu.  Median    Mean 3rd Qu.    Max.
##   431.0   489.0   500.0   499.9   511.0   565.0
```

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



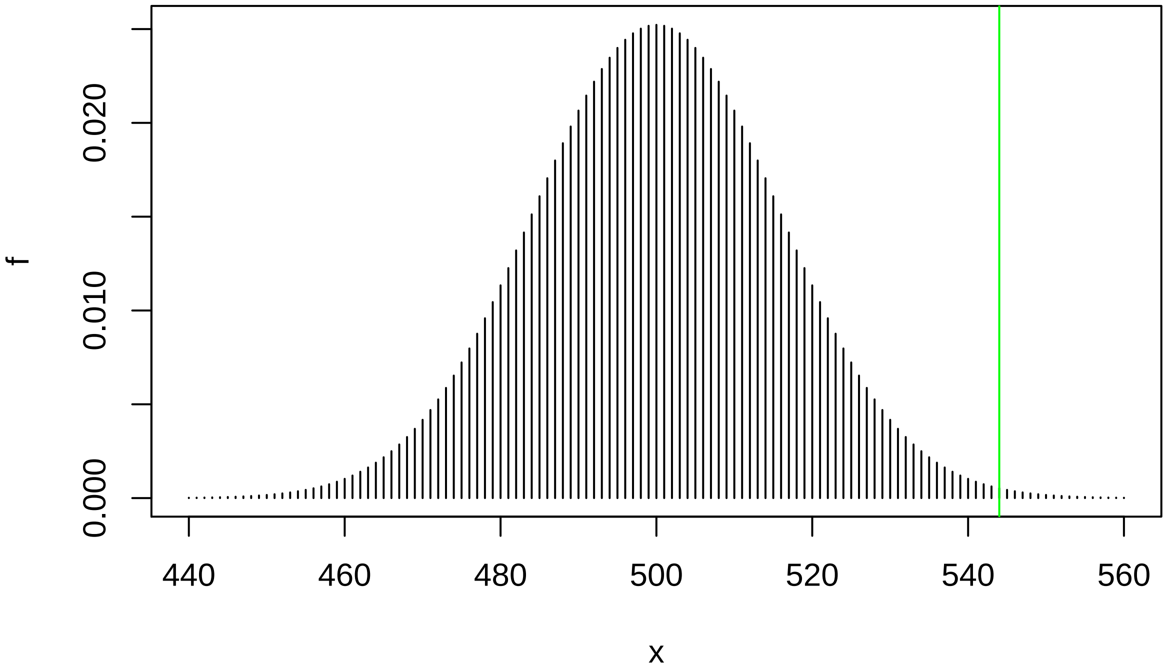
Or we could do exact calculations using the binomial distribution:

```
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?

```
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:

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

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

But if 44 or more 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:

```
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.

### Formal significance testing: The steps

1. Write down the **null hypothesis** ($H_0$) and the **alternative hypothesis** ($H_1$).
2. Choose a **test statistic** that you can find the distribution for when the null hypothesis is true (plus any other necessary assumptions.) Write down the **sampling distribution** for the test statistic, assuming the null hypothesis model is true.
3. Collect random data. Use the observed data to calculate the *observed* value of 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**

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 $\alpha$).

5a. If the $P$-value turns out to be less than $\alpha$, reject the null hypothesis in favor of the alternative. If the $P$-value turns out to be greater than $\alpha$, 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 this section we'll assuming we're doing a test about a population mean $\mu$, so the null hypothesis will be a statement about $\mu$.

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

$$\begin{gathered} H_0 : \mu = 10 \\ H_1 : \mu \neq 10 \end{gathered}$$

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:**

$$\begin{gathered} H_0 : \mu \leq 25 \\ H_1 : \mu > 25 \end{gathered}$$

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

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

$$\begin{gathered} H_0 : \mu \geq 60 \\ H_1 : \mu < 60 \end{gathered}$$

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

To decide what kind of hypotheses to test, first ask yourself: does the test specify a direction? If not, do a two-tailed test. In a two-tailed test for the mean, the null is always "$\mu$ equals something."

If the test specifies a direction, then *the alternative hypothesis is what you're trying to show beyond reasonable doubt.* What you're trying to show will, of course, depend on the application. In academia, the null hypothesis is often "nothing interesting is happening" and the alternative is "something interesting is happening." In a trial of a drug, the null is usually "the drug does no better than a placebo" while the alternative is "the drug does better than a placebo." In business, the alternative might be "this product does better than the status quo." Note that better doesn't necessarily mean higher numbers: a drug that treats high blood pressure should *lower* blood pressure.

## 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 less 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.

A small $P$-value means it would be unlikely to see the data you if the null hypothesis were true. Therefore a small $P$-value means there's evidence against the null hypothesis and for the alternative.

How small is small? Arbitrary rules are arbitrary, but they can provide a good starting point for people new to hypothesis testing. We therefore propose the following rules:

- A $P$-value greater than 0.1 is not small, and provides no real evidence against the null hypothesis.
- A $P$-value less than 0.01 is small, and in most circumstances provides evidence against the null hypothesis and for the alternative.
- A $P$-value between 0.01 and 0.1 is indeterminate, and you have no choice but to think carefully about what the number means in context. Sorry.

Again, these rules are artificial, and hopefully after doing tests for ten years you'll automatically think about what the number means in context.

### 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 $\mu$ 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.

Now, if you do need to make a binary decision, then comparing your $P$-value to a significance level $\alpha$ is fine. But you should choose $\alpha$ based on the context of your problem. Here, $\alpha$ is most easily interpreted as the maximum probability you're willing to accept for wrongly rejecting the null hypothesis. If rejecting the null hypothesis means you'll pause cookie production in your cookie factory to check your cookies are the right size, then a relatively high $\alpha$ like 0.05 might be fine – the cost of a false positive isn't high. If rejecting the null hypothesis means you'll send a person to jail for 50 years, then $\alpha$ should be much lower – you don't want a chance of convincing an innocent person anywhere near as high as 5%.

## 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?

**Step 1: Hypotheses.**

The parameter here is $p$, the probability a Most Beautiful Person's child is a girl. What we would like to show beyond reasonable doubt is $p$ is higher than 0.485, the proportion of girls for the whole population. That'll be the alternative. Everything else goes into the null.

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

Note that some sources will write $p = 0.485$ for the null; mathematically, this would make no difference.

**Step 2: Null distribution.**

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. Let $Y$ be the number of girls of 329 births to the Beautiful People. If the null hypothesis is true, then $Y$ has a binomial distribution with $n = 329$ and $p = 0.485$. (Note that instead of the binomial, we could have used the Central Limit Theorem and the normal distribution – see below. The result would be similar but not quite as accurate.)

**Step 3: Observed test statistic.**

Ideally it's only at this stage that we'd get the data – this makes it much harder to manufacture the result you want. Well, we already have the result: the observed value of the test statistic is 157 girls. We'll have to trust ourselves not to cheat.

**Step 4: Find the P-value.**

Because the alternative is a "greater than," the $P$-value will be the "greater than or equal to" probability. We can calculate this using `pbinom()`:

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

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

**Step 5: What can we conclude?**

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.

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

**Step 6: Find a confidence interval.**

So what might the probability of a Most Beautiful Person's birth being a girl plausibly be? Here's an (approximate) 95% confidence interval:

```
x_bar = 156 / 329
# Lower bound
x_bar - qnorm(0.975) * sqrt(x_bar * (1 - x_bar) / 329)
```

```
## [1] 0.4202082
```

```
# Upper bound
x_bar + qnorm(0.975) * sqrt(x_bar * (1 - x_bar) / 329)
```

```
## [1] 0.5281201
```

The probability of a girl might be anywhere from 42% and 53%. We'd need more data if we wanted to know the probability more accurately.

*Pedant's corner:* We've made some assumptions here. (1) Every Most Beautiful Person's birth has the same probability of being a girl. (2) Whether or not a birth is a girl is independent of every other birth's gender.

These assumptions aren't literally true. (2) isn't true because, for example, Julia Roberts had twins and twins are more likely to be same-sex (because they might be identical), but this probably makes a negligible difference. (1) assumes that every Most Beautiful Person has the same probability of a girl, but this probability migth vary (slightly) from person to person. This is arguably a bigger problem for the confidence interval than the significance test. In a significance test, you're really testing a null hypothesis *model*, and if you reject it you're showing that the null hypothesis model is wrong – maybe because $p$ doesn't take it null value, but also maybe because you made an incorrect assumption. With the confidence interval you're relying on your assumptions being close enough to true that any violations don't matter, which is usually something of an article of faith.

## A test based on the Central Limit Theorem

We know that $\bar{X}$ has an approximately Normal$(\mu, \sigma^2/n)$ distribution for large $n$, so we can create a test based on the Central Limit Theorem when $n$ is large and we have an accurate way of estimating $\sigma$. Suppose the null hypothesis is $H_0 : \mu = \mu_0$. (Here, $\mu$ is the true population mean. $\mu_0$ is our hypothesis for this true population mean: this will just be a number like 20 or 0 or $-2$.)

We usually use a version of test statistic that has a *standard* normal distribution under the null. Let

$$Z = \frac{\bar{X} - \mu_0}{\frac{\sigma}{\sqrt{n}}}$$

If the null hypothesis is true, then $Z$ has an approximately standard normal distribution. The $P$-value will thus be

```
# For a left-tailed test:
pnorm(z)
# For a right-tailed test:
1 - pnorm(z)
# For a two-tailed test:
2 * (1 - pnorm(abs(z)))
# or:
2 * pnorm(-abs(z))
```

The two-tailed case requires some explanation. We need to take the smaller tail probability and double it. Taking `pnorm()` of the *absolute value* of $z$ guarantees this. So does taking one minus `pnorm` of the absolute value of $z$ ensures we get the smaller tail probability. See Trosset p. 214 for a picture.

The usual problem is that we don't know $\sigma$, and the usual solution is to use $s$ instead. We give this test statistic a new name:

$$T = \frac{\bar{X} - \mu_0}{\frac{S}{\sqrt{n}}}$$

For very large $n$, $T$ follows an approximately Normal distribution. $n$ has to be large for two reasons:

- The Central Limit Theorem only implies the distribution gets close to normal for large $n$.
- $S$ has to be close to $\sigma$ with very high probability.

If $n$ is large but not very large, there are adjustments we can make to the distribution to make the test more accurate. If $n$ is very large these adjustments don't make much difference, and we might as well just use the normal distribution because we're used to it.

## Example: Feeling thermometers

In the American National Election Studies (ANES), "feeling thermometers" are used to gauge the public's favorability toward political candidates. A rating of 0 means a person is "very cold" toward a candidate, while a rating 100 means they are "very warm" toward the candidate. A rating of 50 is neutral.

The 2016 ANES pilot study surveyed a sample of voting-age U.S. citizens. 1197 people in the sample gave a feeling thermometer rating for Jeb Bush. The sample mean of these ratings was 34.65 and the sample standard deviation was 26.0.

Can we reject the hypothesis that Jeb's "true" mean feeling thermometer rating (i.e. his average if all voting-age U.S. citizens were asked) was 50?

**Step 1: Hypotheses.**

Let $\mu$ be Jeb's population mean feeling thermometer rating. The way the question was posed doesn't have a direction, so do a two-tailed test.

$$H_0 : \mu = 50$$
$$H_1 : \mu \neq 50$$

**Step 2: Null distribution.**

Our test statistic will be

$$T = \frac{\bar{X} - 50}{\frac{S}{\sqrt{n}}}$$

If Jeb really has a mean feeling thermometer score of 50 and we take a very large simple random sample, then $T$ will have an approximately standard normal distribution. (The ANES data isn't quite a true simple random sample, but we'll overlook this.)

**Step 3: Observed test statistic.**

We now plug in the sample size, sample mean, and sample SD from our actual data.

$$t = \frac{34.65 - 50}{\frac{26}{\sqrt{1197}}} = -20.36$$

If Jeb really did have an average of 50, his mean in the ANES sample would be 20 standard deviations of the error lower than what you'd expect. This isn't looking good for Jeb.

**Step 4: Find the $P$-value.**

The two-tailed $P$-value is

```
t = (34.65 - 50) / (26 / sqrt(1197))
2 * (1 - pnorm(abs(t)))
```

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

The $P$-value is basically zero. If Jeb's population mean feeling thermometer was really 50, it would be virtually impossible to get a sample mean of 34.65 from a sample of size 1197.

**Step 5: What can we conclude?**

If it's virtually impossble to have seen what we saw if $\mu$ really was 50, then the most plausible conclusion is that $\mu$ wasn't 50.

This begs the question – what was $\mu$? Our best guess is that $\mu$ is close to the sample mean – about 35. So Jeb's $\mu$ is almost certainly below 50. But really we want to answer this with a confidence interval.

**Step 6: Find a confidence interval.**

For 95% confidence:

```
x_bar = 34.65
# Lower bound
x_bar - qnorm(0.975) * 26 / sqrt(1197)
```

```
## [1] 33.1771
```

```
# Upper bound
x_bar + qnorm(0.975) * 26 / sqrt(1197)
```

```
## [1] 36.1229
```

Jeb's $\mu$ could plausibly be from 33.2 to 36.1. So not only was Jeb's $\mu$ not 50, it was a long, long way from 50.