Welcome to STA 101!

All of (classical) statistics

Point estimation ("what's the best guess?")

- **Example**: what's my blood pressure?
- **Jargon**: point estimate

Interval estimation ("plus or minus what?")

- **Example**: what's a likely range for my blood pressure?
- **Jargon**: confidence interval, confidence level, ...

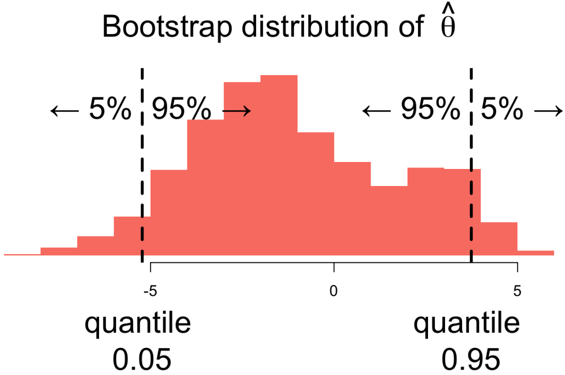
Hypothesis testing ("yeah, great, so...what can we conclude?")

- **Example**: soooo...am I at risk, or not?
- **Jargon**: test statistic, null distribution, significance level, p-value, Type 1 and 2 errors, power, ...

Last time...

In the beginning: we need an interval to capture uncertainty.

In the end: Ta-da! An interval...



In between: confusing and ill-motivated babble.

Interval estimation
in
18 questions

1. **Q**: What is an interval estimate?

   A: a range of plausible values for the answer to a question.

2. **Q**: Why do we want an interval estimate?

   A: To communicate uncertainty about the point estimate. If the range is wide, uncertainty is high. If the range is narrow, uncertainty is low.

3. **Q**: Why do we care?

   A: Important decisions may be made based on the results of a data analysis, but those decisions will look very different depending on the uncertainty that is attached to the results.

   Example: We know a particular gene makes you immune to a certain disease. If 12 people in a sample of 40 have this gene, our best guess is that 30% of humans have the gene, but should we be making bold, confident medical decisions based on this information? Probably not.

4. **Q**: What uncertainty do we seek to capture with our interval?

   A: In this class, primarily *sampling uncertainty*.

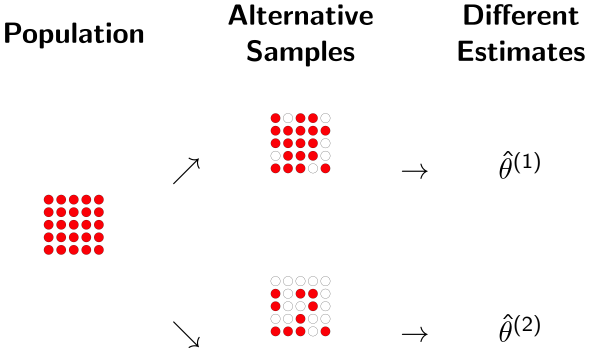
5. **Q**: What is sampling uncertainty?

   A: In practice, we have a single random sample from the population. It is meager and imperfect. Nevertheless, we *hope* that the sample resembles the population closely enough that we get accurate estimates. But we are uncertain of this.

6. **Q**: Why are we worried about this?

   A: An *unrepresentative* sample can lead to a distorted, misleading, and inaccurate estimate. If I knew that that's what I had, I would throw it in the trash. But how do I know?

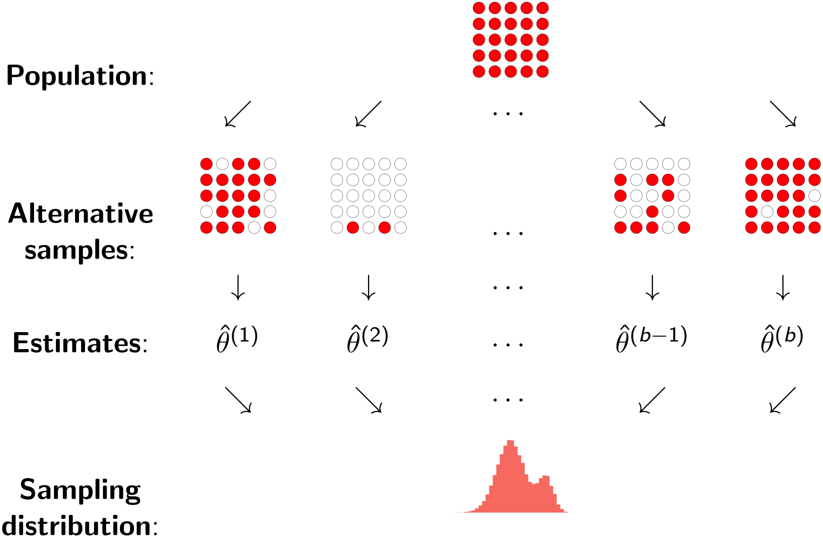
7. **Q**: What if we had a different sample to compare to?



- Different sample $\implies$ different answer. But how different?
- **Reassuring**: I collect a whole other sample and it gives me basically the same thing. Phew!
- **Alarming**: I collect a whole other sample and it gives me a totally different answer. Yikes!

8. **Q**: How do we make this comparison precise?

- We look at the *sampling distribution* of the estimates;
- Different sample $\implies$ different estimates;
- Instead of comparing just two different samples, we imagine comparing 100, or 1000, or 10^6;
- The sampling distribution visualizes the variation in those different estimates.

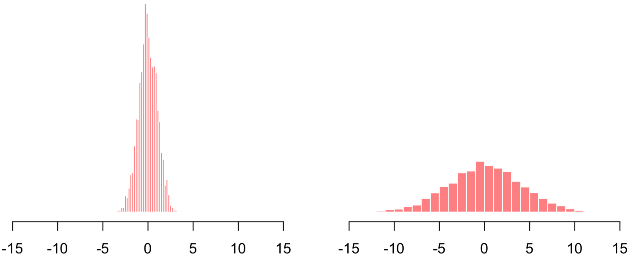


Population:
Alternative samples:
Estimates:
$\hat{\theta}^{(1)}$
$\hat{\theta}^{(2)}$
$\hat{\theta}^{(b-1)}$
$\hat{\theta}^{(b)}$
Sampling distribution:

9. **Q**: What does the sampling distribution tell us?

Different sample $\implies$ different estimates.

How worried should I be about this? Look at the spread:



- **left**: sampling uncertainty is lower, and I'm less worried;
- **right**: sampling uncertainty is higher, and I'm more worried.

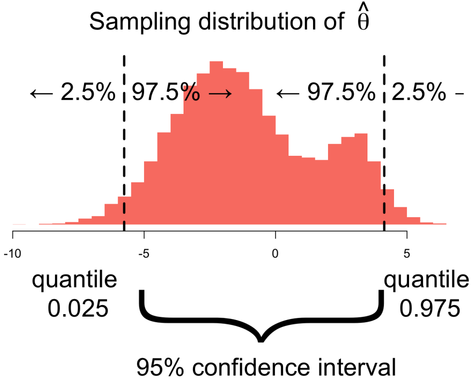
The spread tells you how sensitive the estimate is to the sampling. This sensitivity should be low if our results are truly reliable.

10. **Q**: What does this have to do with interval estimation?

    A: We said we wanted an interval that captures sampling uncertainty. The sampling distribution displays *all* the sampling uncertainty. So if we pick a range of values that includes "most" of the sampling probability, we're done.

11. **Q**: How do you do that?

    A: There are many ways, but the easiest is to use quantiles:



12. **Q**: Wait a minute. In order to actually pull this off, I need to be able to create many random samples from the population. Can we actually do this?

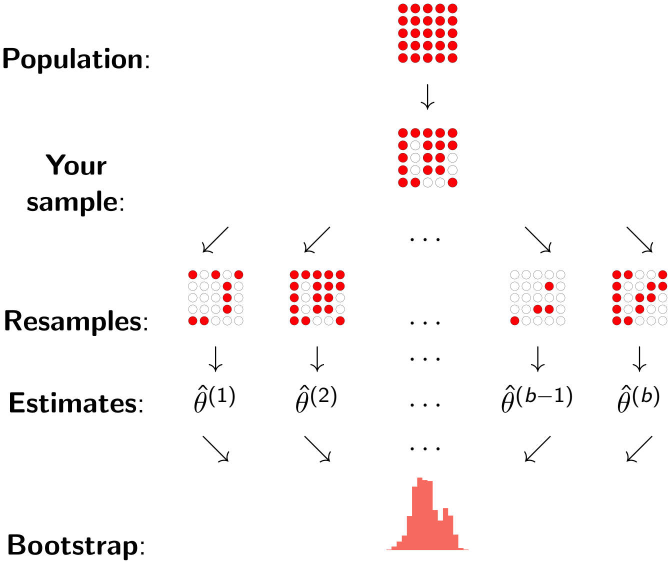
    A: In general, no. It is too costly. And besides, if you could freely generate more data at will, wouldn't you have done it already to create a bigger, more complete dataset?

13. **Q**: Right...so what do I do?

    A: You *approximate* the sampling distribution with something called the bootstrap.

14. **Q**: WTF is the bootstrap?

    A: It is a leap of faith. Pretend that your original sample *is* the population, and generate your alternative, hypothetical samples from that. It's messy, and it might not always work, but it's the best you can do.

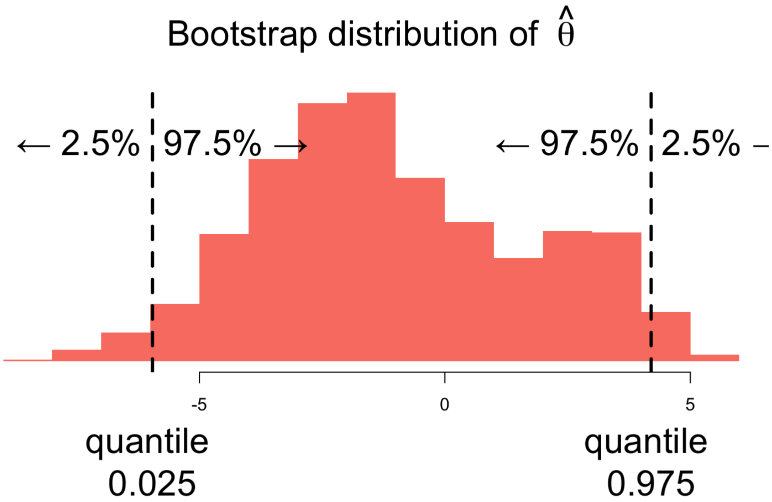

Population:
Your sample:
Resamples:
Estimates:
$\hat{\theta}^{(1)}$ $\hat{\theta}^{(2)}$ . . . $\hat{\theta}^{(b-1)}$ $\hat{\theta}^{(b)}$
Bootstrap:

15. **Q**: What does it mean to generate a new dataset from an old one?

    A: You fill in the rows of a new data frame by randomly picking rows from your old one *with replacement*. Some of the original rows will not appear at all. Some will be repeated. But it's a new data frame, and you can calculate an estimate with it and see if that estimate is close or far from what you originally had.

16. **Q**: And this will finally give me an interval that captures sampling uncertainty?

    A: Approximately. If the leap of faith doesn't kill you first.



Bootstrap distribution of θ̂
← 2.5%
97.5% →
← 97.5%
2.5% →
-5
0
5
quantile
0.025
quantile
0.975

17. **Q**: How do you decide the confidence level of the interval? 75% vs. 90% vs. 95% vs. 99% etc?

    A: It will depend on context, But in general you face a trade-off. You want an interval that is...

    - ...large enough to capture the truth;
    - ...small enough that it actually teaches you something meaningful about where the truth is likely to live.

    The interval $(-\infty, \infty)$, for example, is guaranteed to include the truth 100% of the time, but it is completely uninformative. Adjusting the confidence level is one way to negotiate this trade-off.

18. **Q**: When we compute a 95% confidence interval (or 90% or 99% or whatever), how doweyou interpret the 95% part?

    A: 95% refers to the reliability *in repeated use* of the method that generated the interval. It does not refer to the reliability of a particular interval in a single use.

    Imagine every day we collect a new data set (a new random sample from the same population), and we compute a new 95% interval based on it. We repeat this process for 100 days. We are saying that, on 95 of those days, the interval we computed will contain the truth. *On any given day*, the truth is either in the interval or it is not, and for a particular interval we won't be able to tell if we got lucky or unlucky. But 95% of the time, in repeated use, the interval contains the truth.