STAT0032: Introduction to Statistical Data Science

Dr. Francois-Xavier Briol

Department of Statistical Science
University College London

Week 3:
Confidence Intervals

Introduction to Confidence Intervals

Recall our earlier NHANES data from Week 1

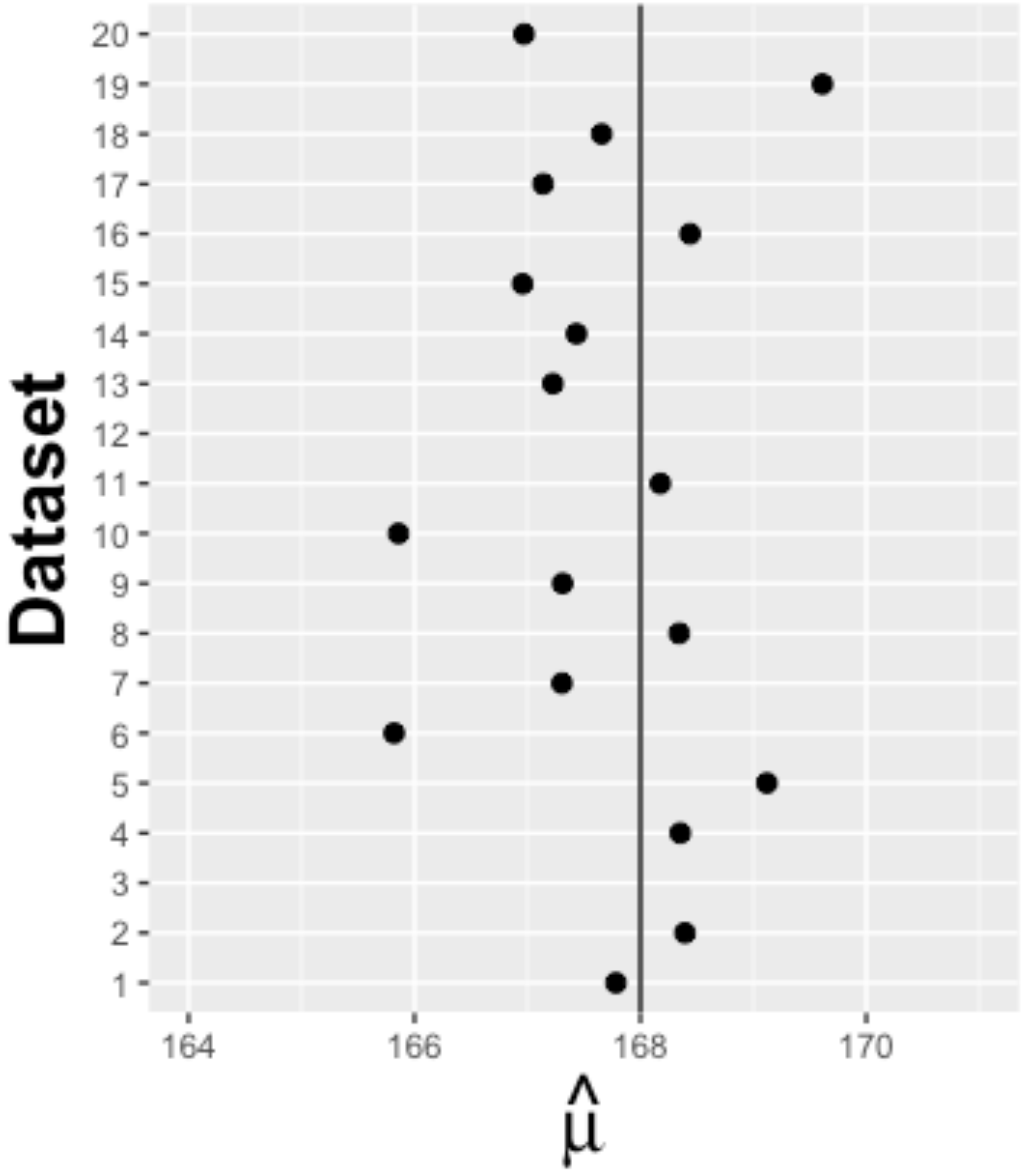
- We looked at the distribution of heights within the data set and estimated the expected height to be 168cm.
  - $\hat{\mu} = 168.$

- We also asked the question: what would have happened if we had sampled a **different** set of 19 219 individuals **from the same population**?

Using simulation to understand confidence intervals

- Imagine the following experiment: we generate **simulated data** like we have been in previous R examples.
- This is called **(Monte Carlo) simulation**, for which there are algorithms based upon **pseudo-randomness**.
- We generate datasets of size 50, each from a distribution.
  - We repeat this process multiple, say 20, times.
  - Take a look at the prepared R demonstration.

Simulation Results (I)



Using simulation to understand confidence intervals

- The sample average is of course a random variable, since it is a function of the data.
  - $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$

- Importantly, the sample average is a statistic – a function of only the data and not of the unknown model parameters.

- However, the *distribution* of the sample average should be a function of the data distribution.

Recall from Week 1

- If

$$X \sim N(\mu, \sigma^2).$$

- Then

$$E[X] = \mu.$$

- You could (re-)check this if you wished.

$$E[X] = \int x \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}\right\} dx = \mu$$

Estimator of the Mean

- $$E[\bar{X}] = E\left[\frac{1}{n}\sum_{i=1}^{n} X_i\right] = \mu$$
  - This definition follows from the general definition of the expectation of a random variable.

- For this reason we may consider $\bar{X}$ to be a plausible estimator of μ.
  - We typically denote estimators with hats and so in this case we choose $\hat{\mu} = \bar{X}$.

More than this: Transformations of Gaussians

- We may characterise the whole distribution of the sample average if each $X_i \sim N(\mu, \sigma^2)$:

$$\bar{X} \sim N\left(\mu, \frac{\sigma^2}{n}\right).$$

- We can use this to our advantage.
  - Consider the distribution of

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}}.$$

Pivot

- We may use the result that $Var\left(\frac{Y}{c}\right) = \frac{1}{c^2}Var(Y)$, for Y any random variable and c any constant, and that linear transformations of normal random variables are themselves normal random variables.

- As a result:

$$\frac{\bar{X} - \mu}{\sqrt{\sigma^2/n}} \sim N(0,1).$$

  - This is a **pivot** – a function of the parameter of interest which has a known distribution.
  - Pivots are typically hard to find.

Bounding

- Now that we know the distribution of this statistic, we can make probabilistic statements such as:

$$P\left(\frac{\bar{X}-\mu}{\sqrt{\sigma^2/n}} \leq 1.5\right) = P(Z \leq 1.5) \approx 0.93.$$

  - The choice of 1.5 as a cutoff here is arbitrary and you could consider any alternative.

- Using this result, we may claim that

$$P\left(\mu \geq \bar{X} - \frac{1.5\sigma}{\sqrt{n}}\right) \approx 0.93.$$

Two important observations

- In general, $\bar{X} - 1.5\sigma/\sqrt{n}$ is not a statistic.
- Assume to begin with that σ^2 is known, to simplify the argument.
  - In that case, the above is a statistic.
- But, how do we interpret the statement:

$$P(\mu \geq \bar{X} - 1.5\sigma/\sqrt{n}) \approx 0.93$$

The key point concerning confidence intervals

- The randomness which we are quantifying using the previous statement is not on the parameter μ.

- The randomness is in the data, which is summarised in this example by $\bar{X}$.

- So what does this mean?
  - After all, we have the data right in front of us.

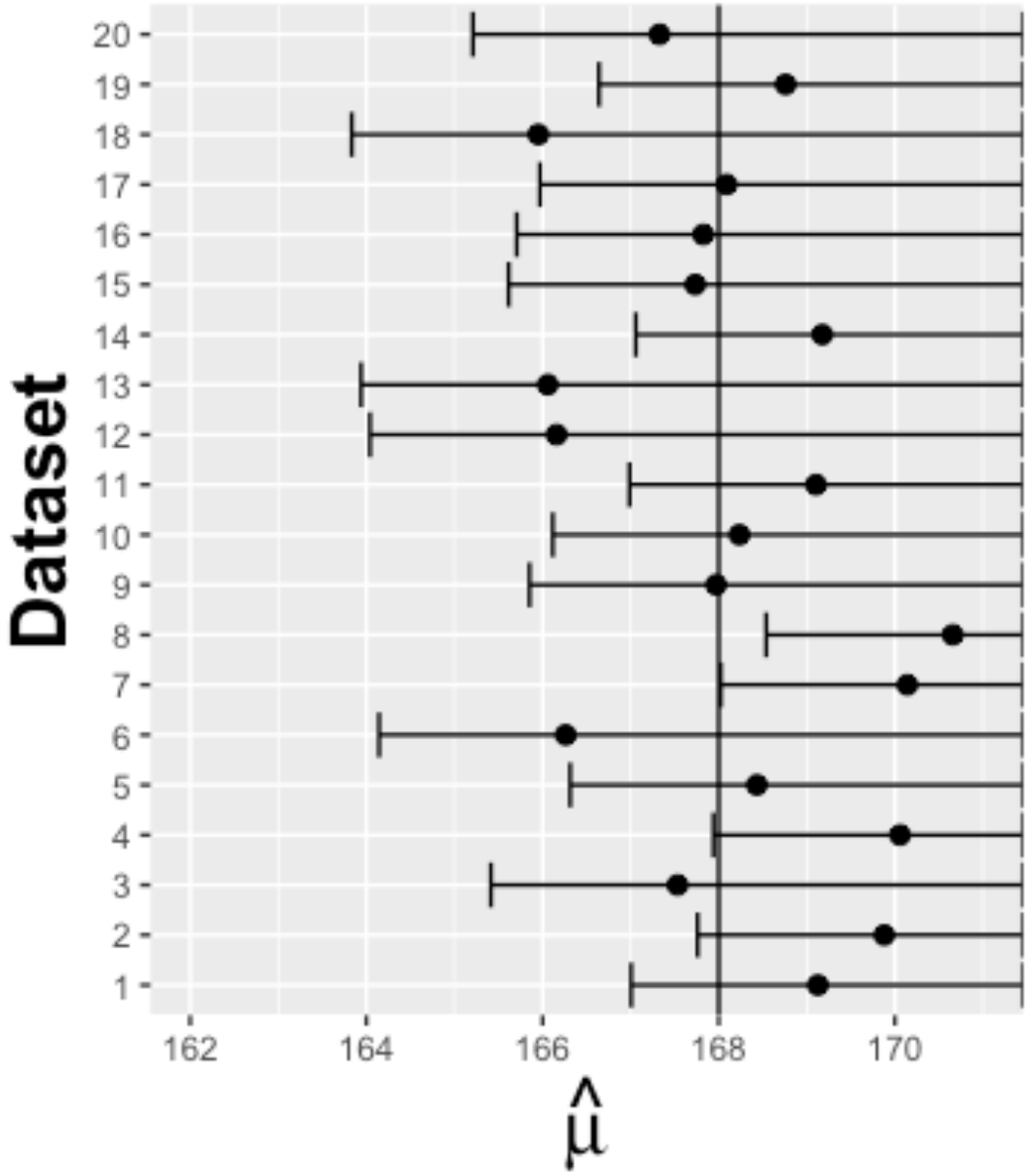
Coverage

- Regardless of what the value of μ is, if the sample size is n and the data follows a $N(\mu, \sigma^2)$ distribution then, in the limit of infinite repetitions of the dataset, the interval $[\bar{X} - 1.5\sigma/\sqrt{n}, \infty)$ will contain μ (approximately) 93% of the time.

- Another way of stating this is that the **coverage** of the interval is 93%.

- We may examine this result using an R demonstration.

Simulation Results (II)



Lessons

- As previously mentioned, no-one expects you to collect “infinitely many data sets” for the same problem.
- Instead, if you provide a (say) 93% confidence interval for your quantity of interest in every analysis you work on throughout your career, then in the long run those provided intervals will include the true quantity of interest 93% of the time.
- **You cannot know (without further data) for which intervals you got it right, just the long run performance!**

In practice

- There will be several assumptions in your model that will be violated, so coverage will not be exact.
- Despite being an idealization, reporting confidence intervals is of major importance in many applications.
  - **At the very least as a way of being more humble about which conclusions you can draw.**
- Just because some modern models have difficult-to-interpret parameters (eg. neural nets), it doesn't mean that confidence intervals will not be used at some point in your application (eg. in the estimation of the empirical performance of the neural net).

In practice

- As we progress through the course and get onto later topics, we will talk about confidence intervals in those contexts.
  - One such topic is regression.

Another interval

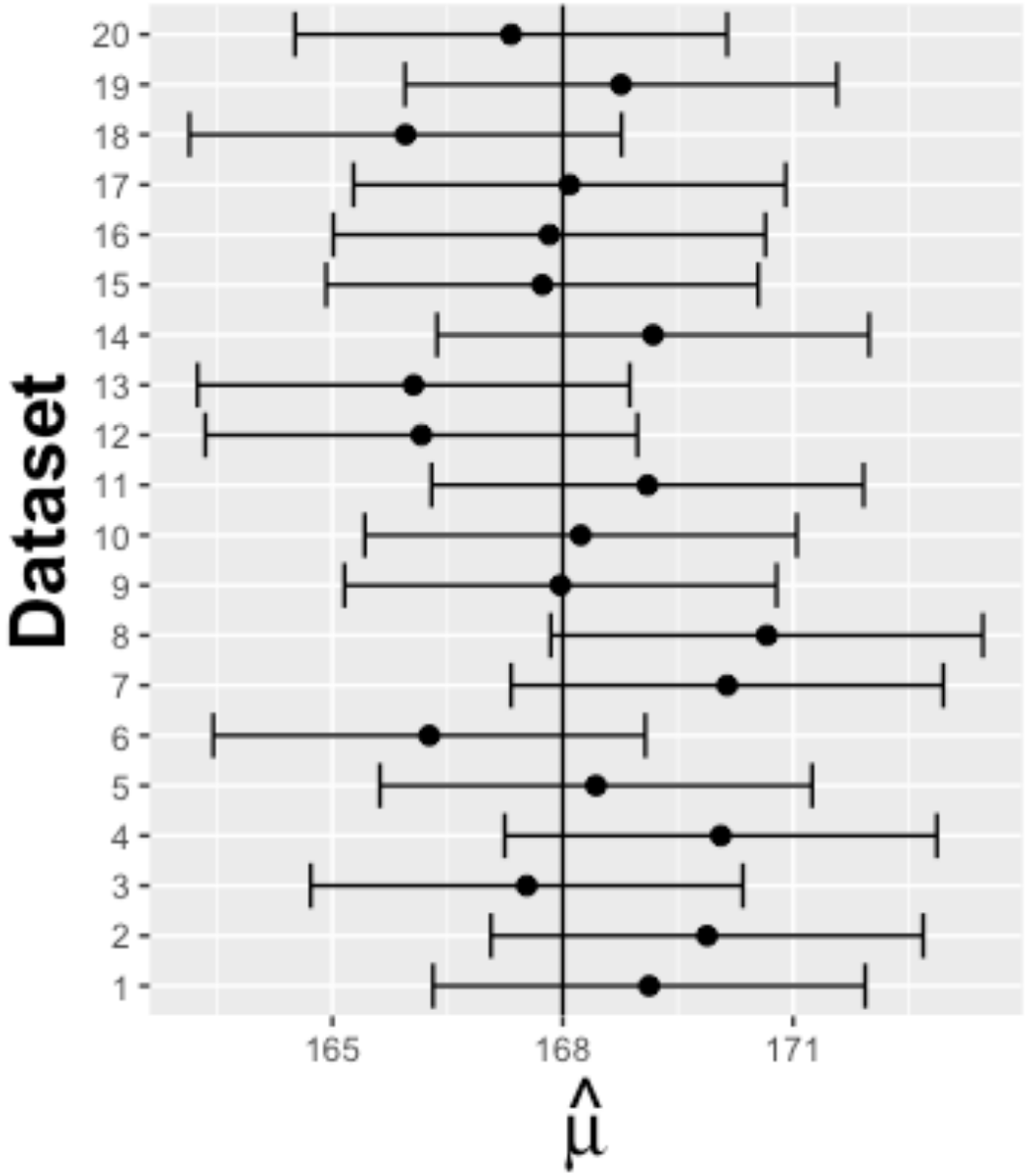
- It is much more common to report lower and upper bounds to create a two-sided confidence interval.
- Returning to the previous example, if you were after a 95% confidence interval (the default in most software) then you can achieve this by determining the 2.5% and 97.5% quantiles of the distribution of the statistic.
  - For instance

$$[\bar{X} + z_{0.025}\sigma/\sqrt{n}, \bar{X} + z_{0.975}\sigma/\sqrt{n}]$$

2.5% quantile of N(0, 1)

97.5% quantile of N(0, 1)

Simulation Results (III)



Confidence Intervals from the Central Limit Theorem

Recall our Two-sided Interval For Gaussian Data

- We have previously seen that if the data is IID from a known Gaussian distribution, then it is possible to obtain the following confidence interval for the mean:

$$[\bar{X} + z_{0.025}\sigma/\sqrt{n}, \bar{X} + z_{0.975}\sigma/\sqrt{n}]$$

- But what if these assumptions are not satisfied? What if we don't know the variance?

A “real” statistic for the Gaussian mean confidence interval

- Previously we proposed a confidence interval for the mean of the normal distribution under the assumption that the variance was known.

- Instead, consider

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \sim \mathcal{T}(n-1), \qquad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2.$$

  - S^2 is the sample variance.

- Exercise: can you use this to derive a 95% confidence interval?

A “real” statistic for the Gaussian mean confidence interval

- Notice: for large n, which we are assuming anyway, the following is used in practice

$$\frac{\bar{X} - \mu}{\sqrt{S^2/n}} \approx N(0,1).$$

- The use of ≈ in this expression is made to indicate that the distribution is approximate and not exact.

Those bounds again

- Suppose you want to trap a parameter of interest, θ, with coverage probability c.
  - A general template is then the confidence interval
$$P\left(lower_c(X) \leq \theta \leq upper_c(X)\right) = c.$$
  - So, depending upon the choice of c, you will obtain (random) upper and lower bounds which depend upon the data X

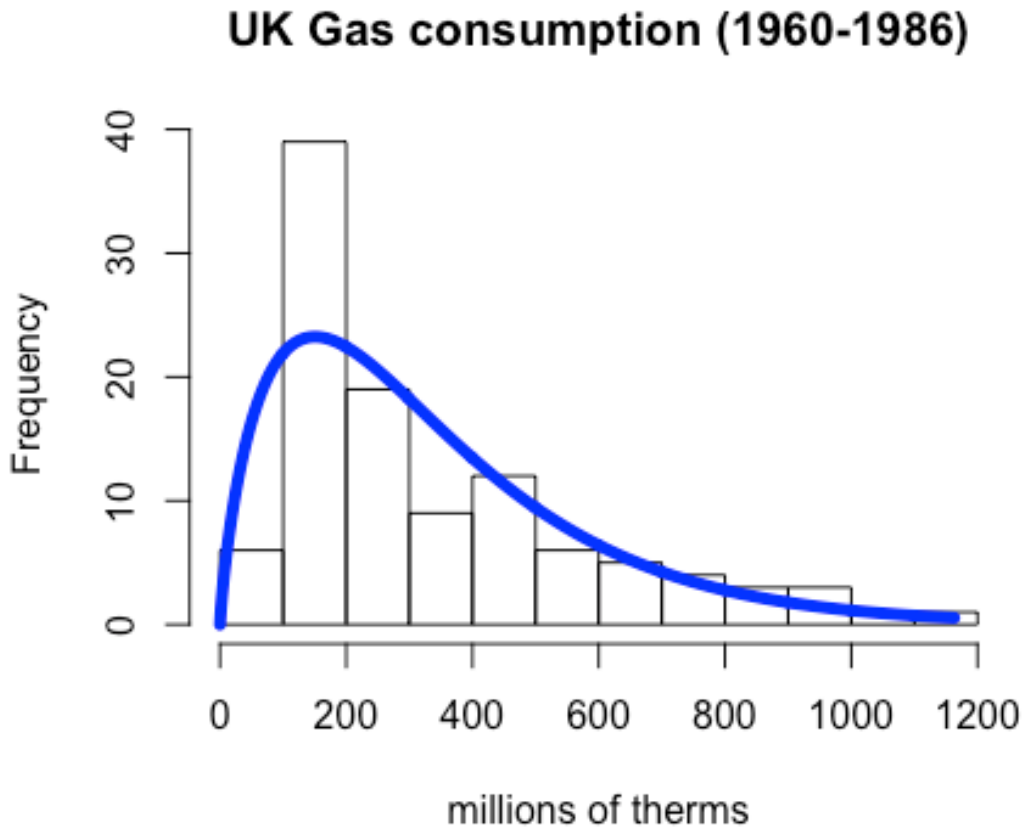
- In general, it is not at all easy to find the actual distribution of the upper and lower bounds.
  - It would be crazy to assume Gaussainity in general, however, the normal distribution is useful in a different way.

The Central Limit Theorem to the rescue

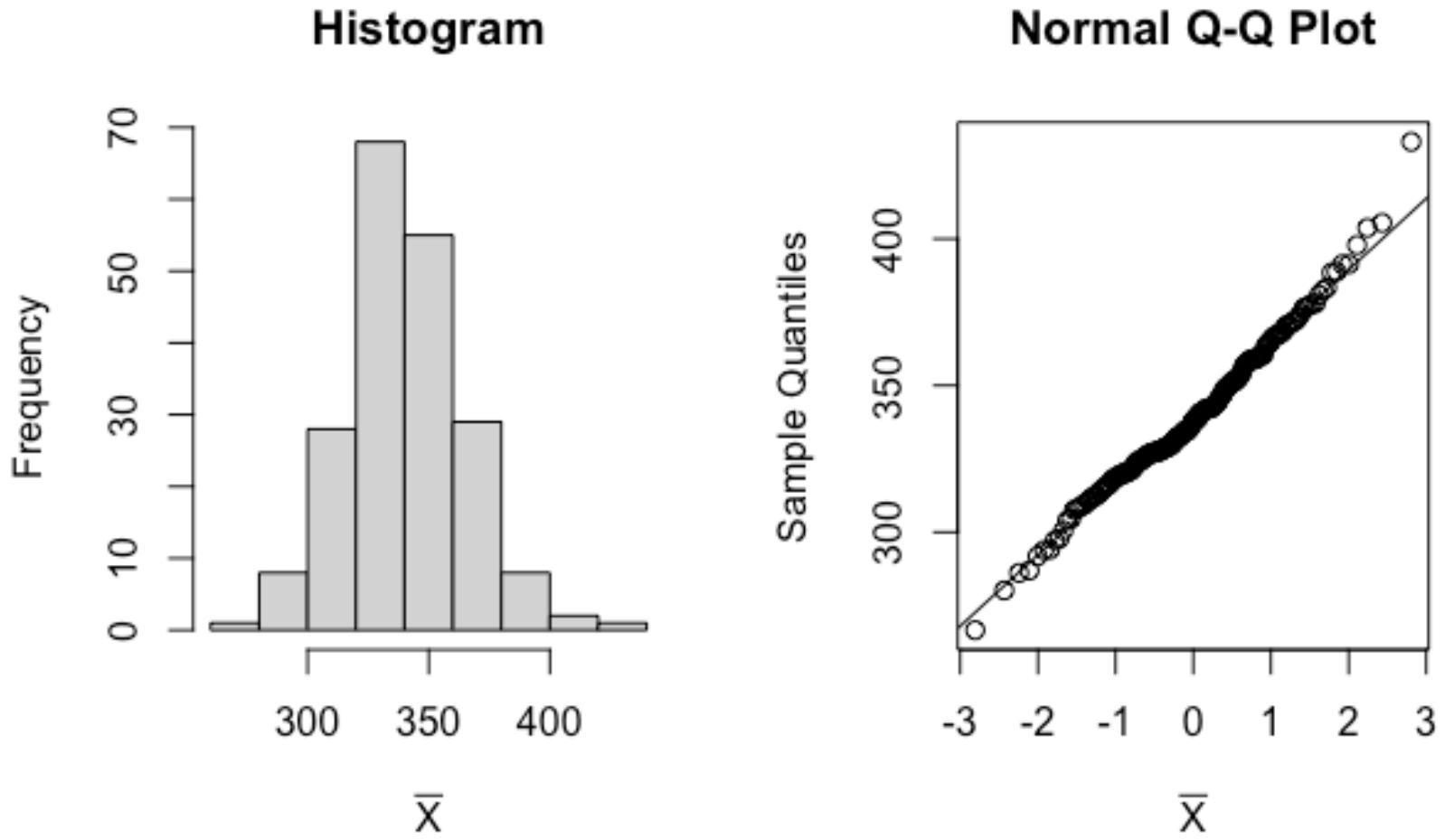
- Many of the upper and lower bounds introduced on the previous slide depend upon averages.
- We have seen that the distribution of the mean of a large number of samples converges to the normal distribution.
  - By the Central Limit Theorem.
- Take a look at the corresponding R demonstration to support the CLT.

UK Gas Consumption Example

- We fit a Gamma distribution to historical data of UK Gas comsumption (108 data points).
- We then resample 200 times 108 data points from this distributions and compute the sample mean.



UK Gas Consumption Example: Some Results



Practical advice

- Many confidence intervals in software packages rely upon the Central Limit Theorem under the hood.
- The Central Limit Theorem makes a statement about large sample sizes.
  - How large is large enough?
  - The subfield of statistical asymptotics is the place where research into this happens.
  - As a result of such work, there are sometimes checks which can be carried out.

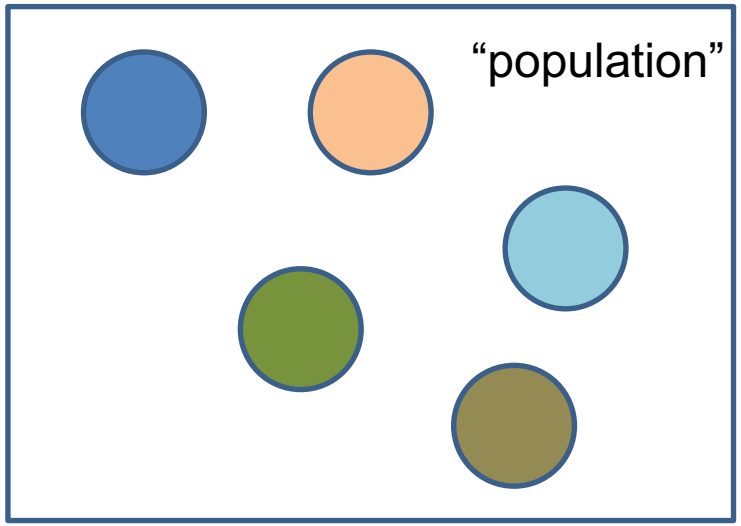
Resampling and the Bootstrap

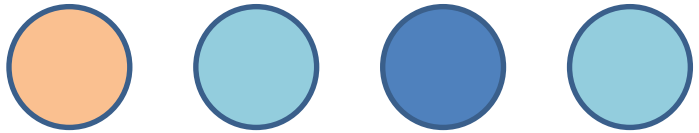
A missing ingredient

- Normal approximations for averages are fine in many cases.
  - But, what is the variance of your statistic?
- In some cases we may write down the variance.
  - In other cases it may be impossible to do so.
- Computer-intensive alternatives have been developed for cases where the algebra fails.
  - Begin, the **bootstrap**.

The bootstrap

- Our data is a (hopefully representative) sample from a population.
- What if we consider our observed data to be the whole population?





“re-samples”

Sampling with replacement and sample size

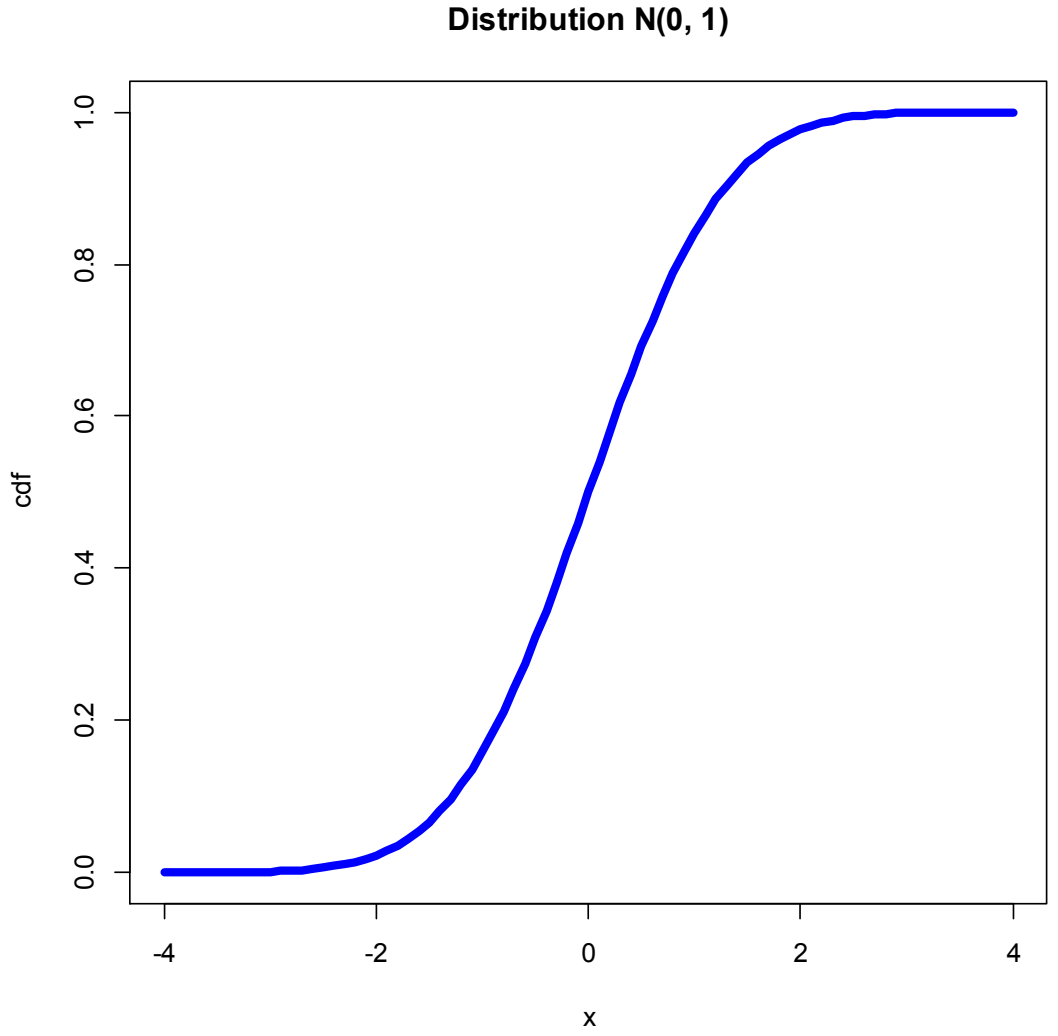
- A foundation of the bootstrap is **sampling with replacement**.
  1. Randomly select a data point.
  2. Add it to the “re-sample”.
  3. Put it back in the box.
- We use sampling with replacement to generate a re-sample of the same size as the original sample.
  - We mimic the process that generated the data, including the sample size.
- We are using the data we have to generate data.
- The purpose of this is to inform us about the sampling variability

Why resample?

- We perform resampling with replacement because we can do it many, many times to simulate in a computer the idea of “infinite replicates of a dataset”.
  - Ideally, we would do infinitely many replicates.
  - In practice, we choose a large number and accept that there will be a degree of approximation error.
- We then quantify how our statistic varies across these synthetic replicates.

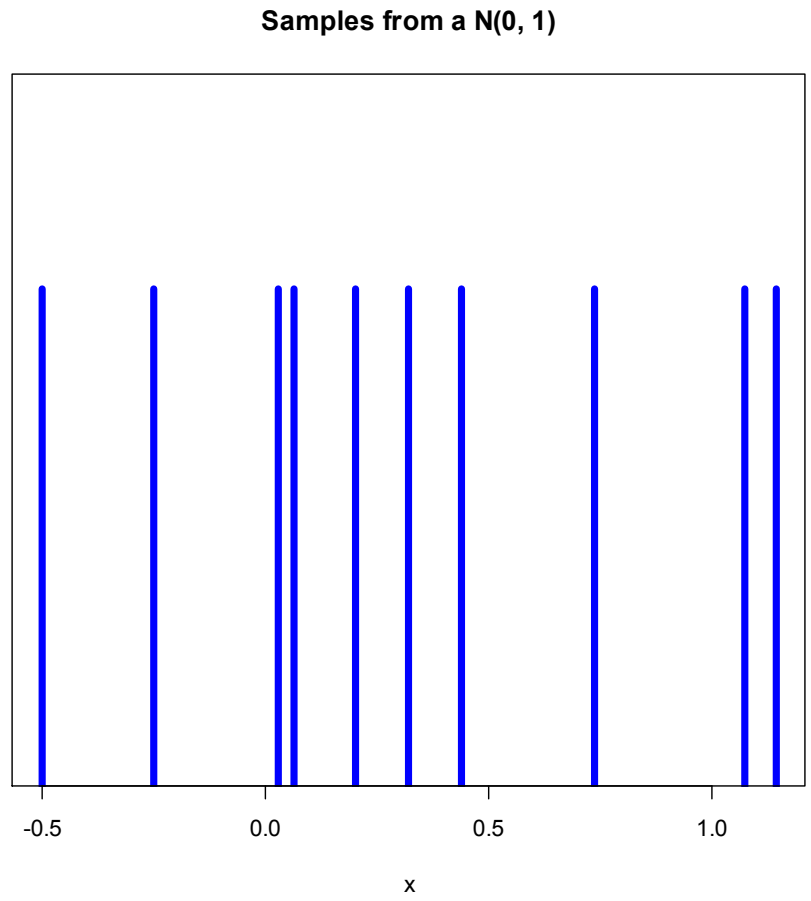
An intuition for why this works

- We have previously seen the Cumulative Distribution Function,
  $F(x) = P(X \leq x)$.



The Empirical Cumulative Distribution Function

- Consider the thought experiment in which our sample is the entire population.
- What is the Cumulative Distribution Function when new data should be only at particular locations, each with equal probability?
- The plot on the right shows 10 draws from N(0,1).

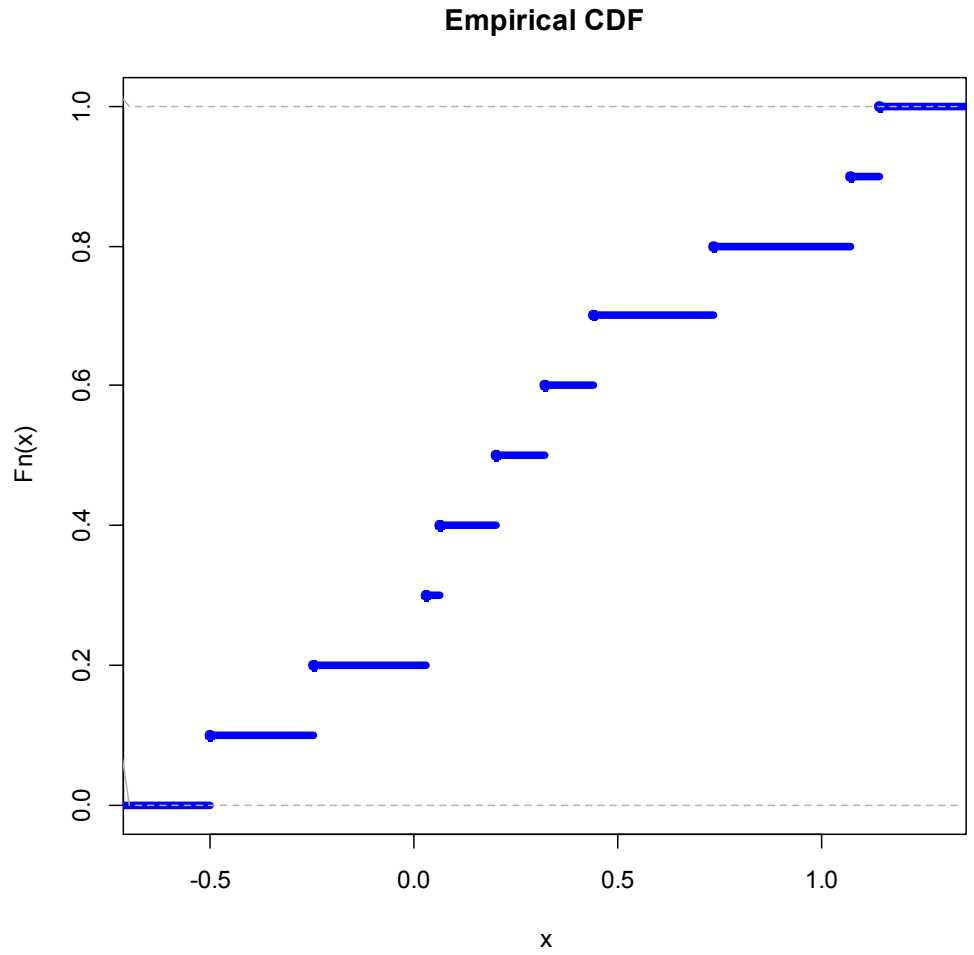


The Empirical CDF

- For any particular level x, we may then determine the empirical CDF to simply be the proportion of points no larger than x.

- $\hat{F}_n(x) = \frac{no.\,points \leq x}{no.\,points}$.

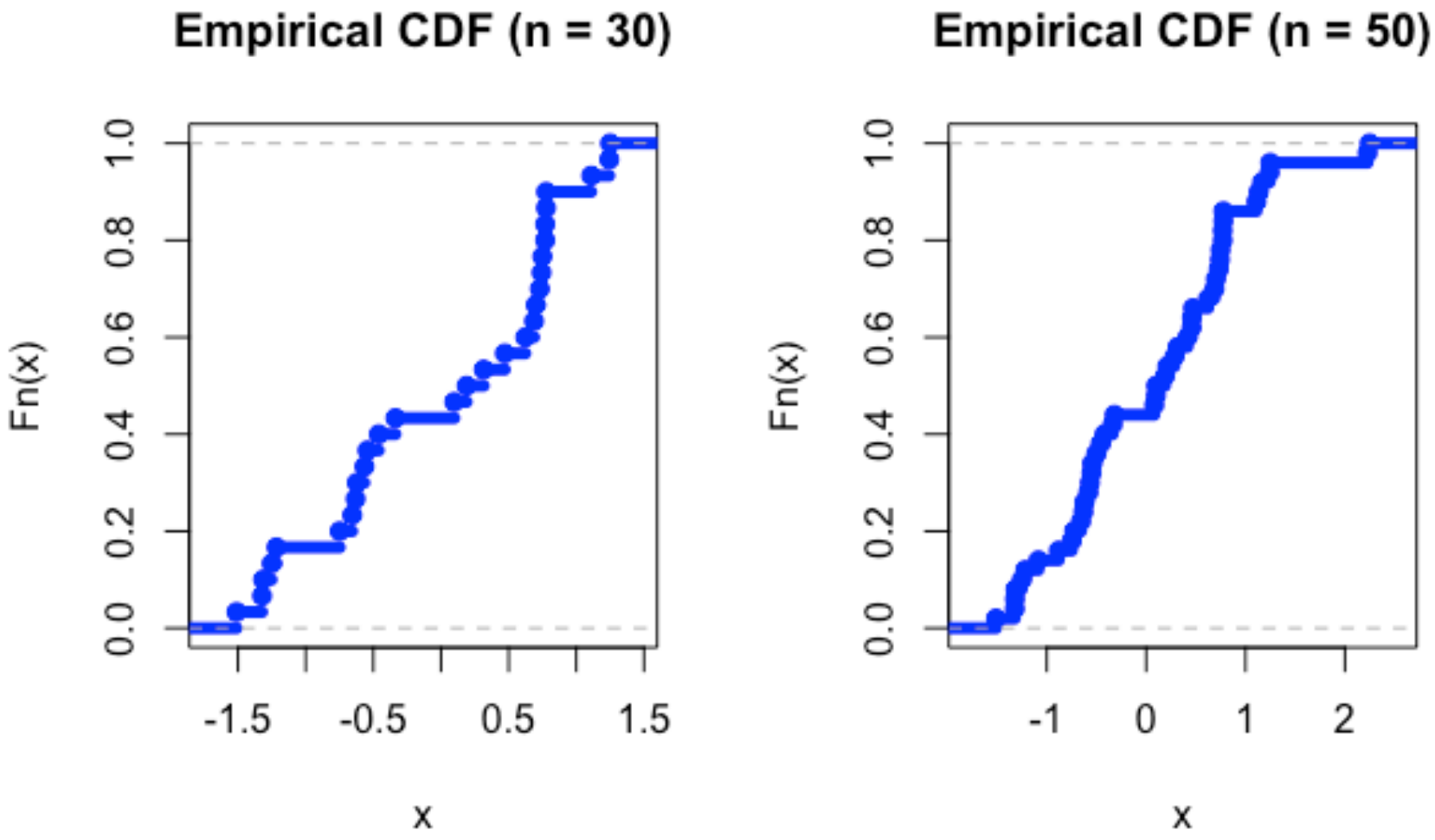
- $\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(x_i \leq x)$.



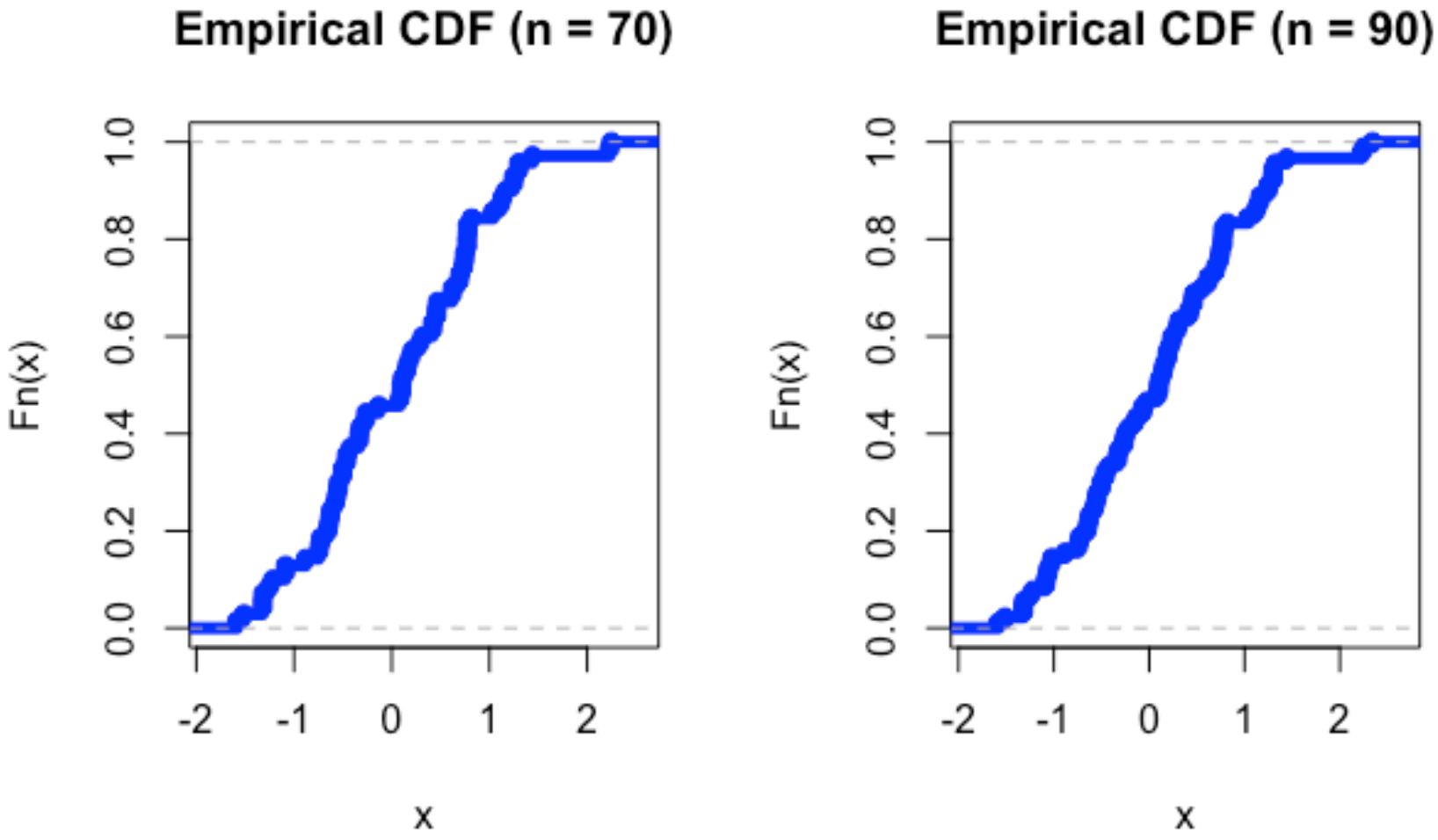
In the limit

- Analogous to the law of large numbers (LLN), empirical CDFs converge to population CDFs.
  - LLN: sample means converge to the expected value of their distribution as the sample size increases.
  - We can visualise this using an R demonstration.
- Informally, we therefore say that the empirical CDF carries some information about sampling under the true CDF.

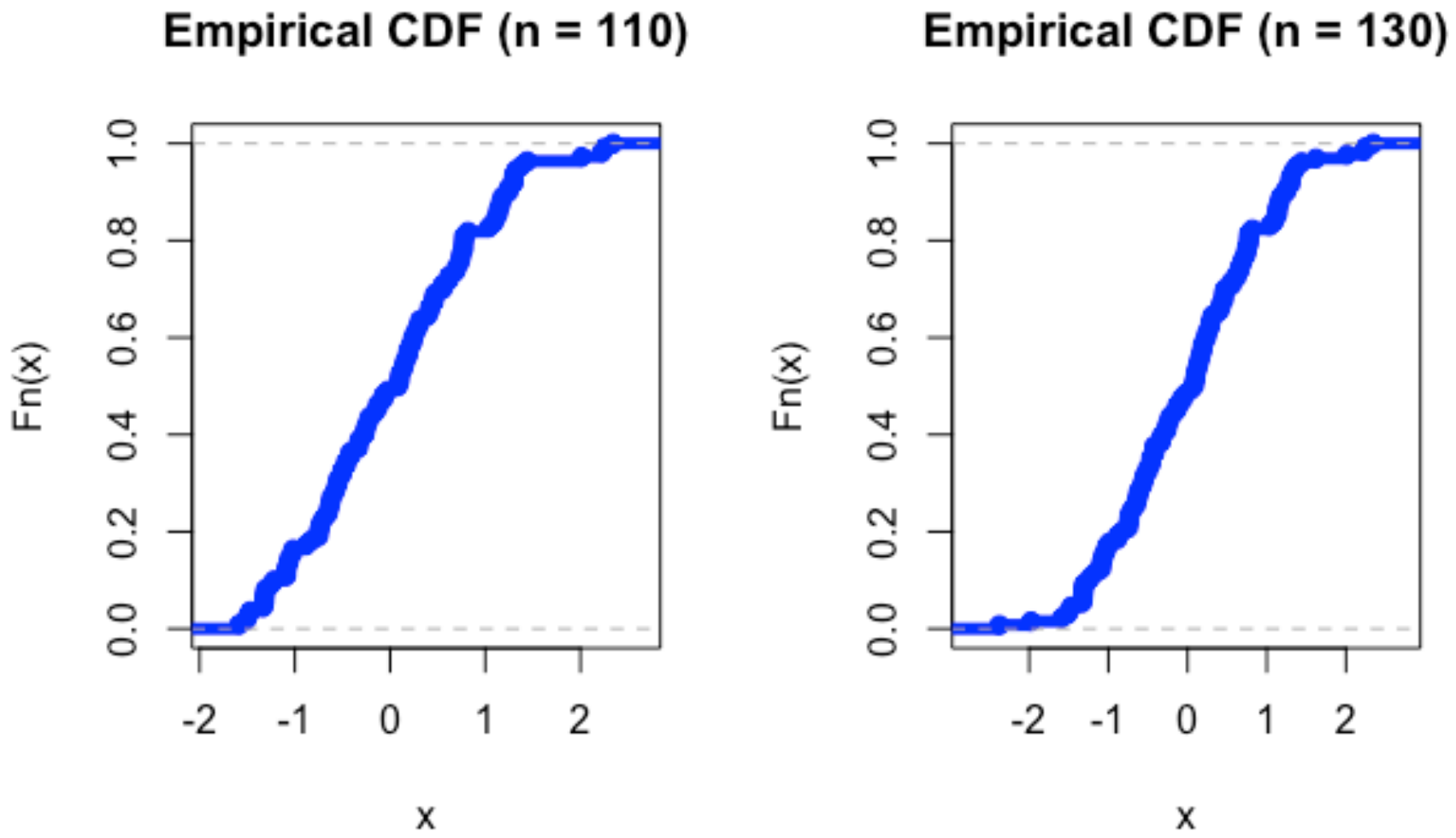
Empirical CDF with n=30 and n=50



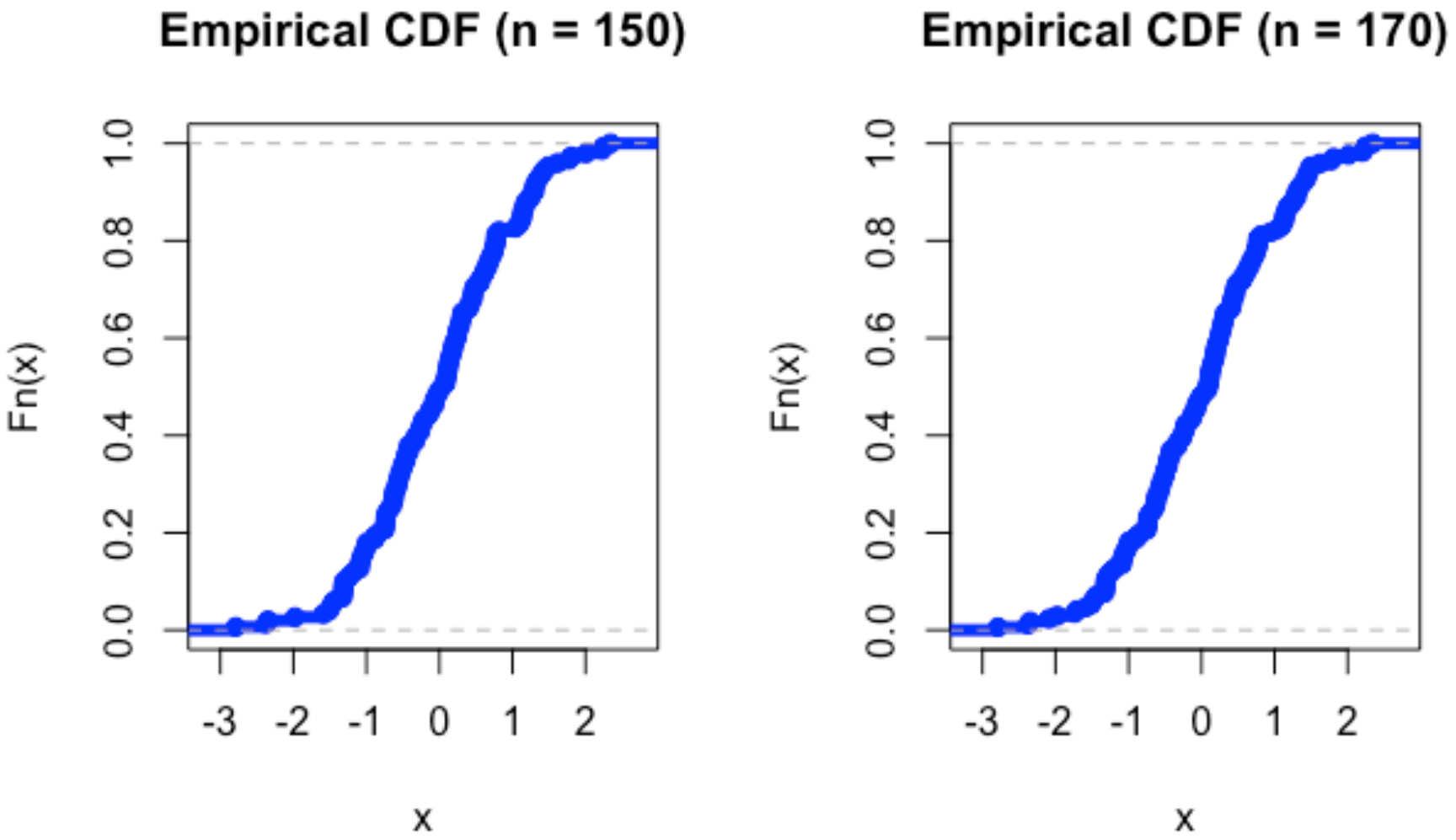
Empirical CDF with n=70 and n=90



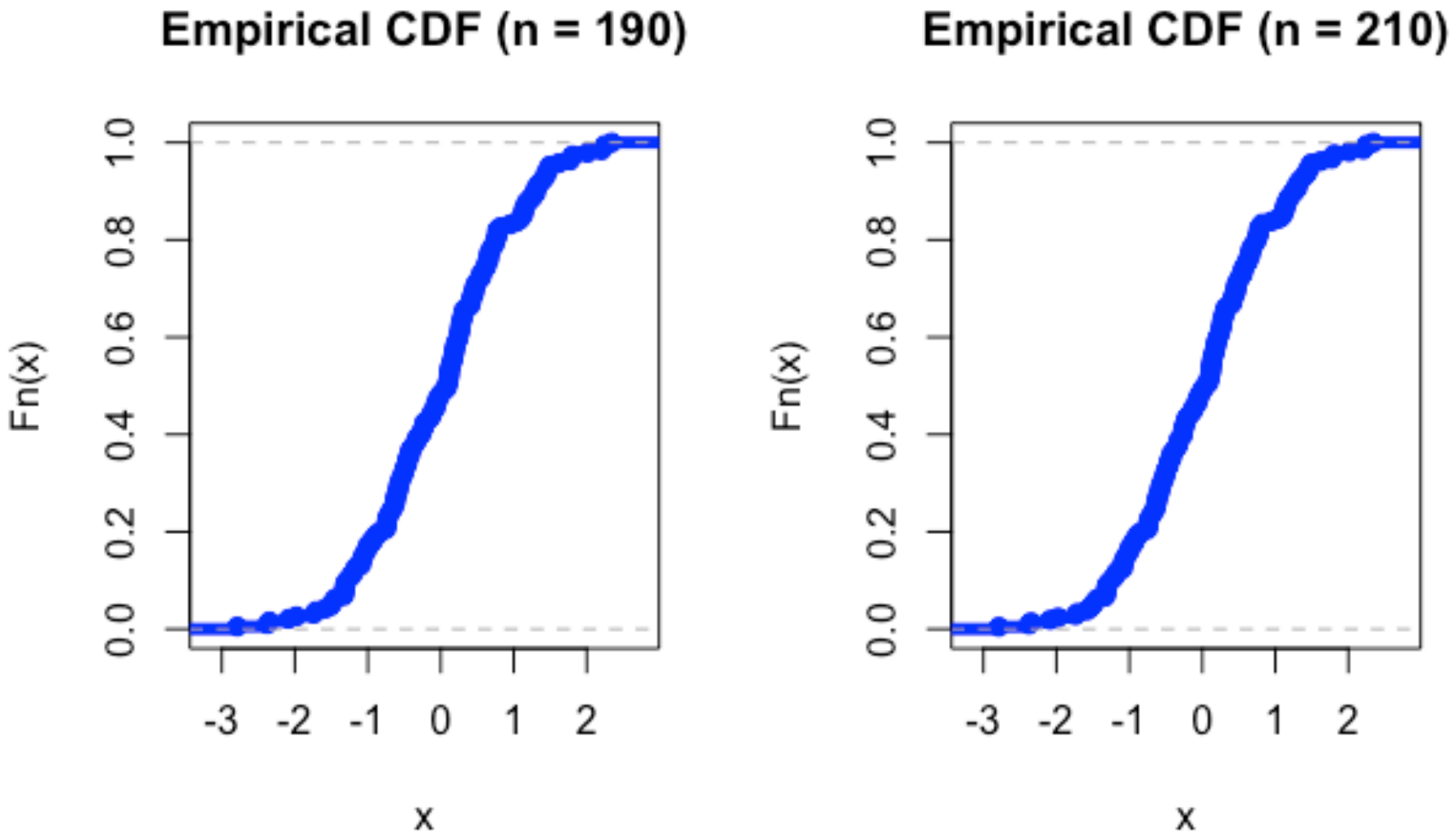
Empirical CDF with n=110 and n=130



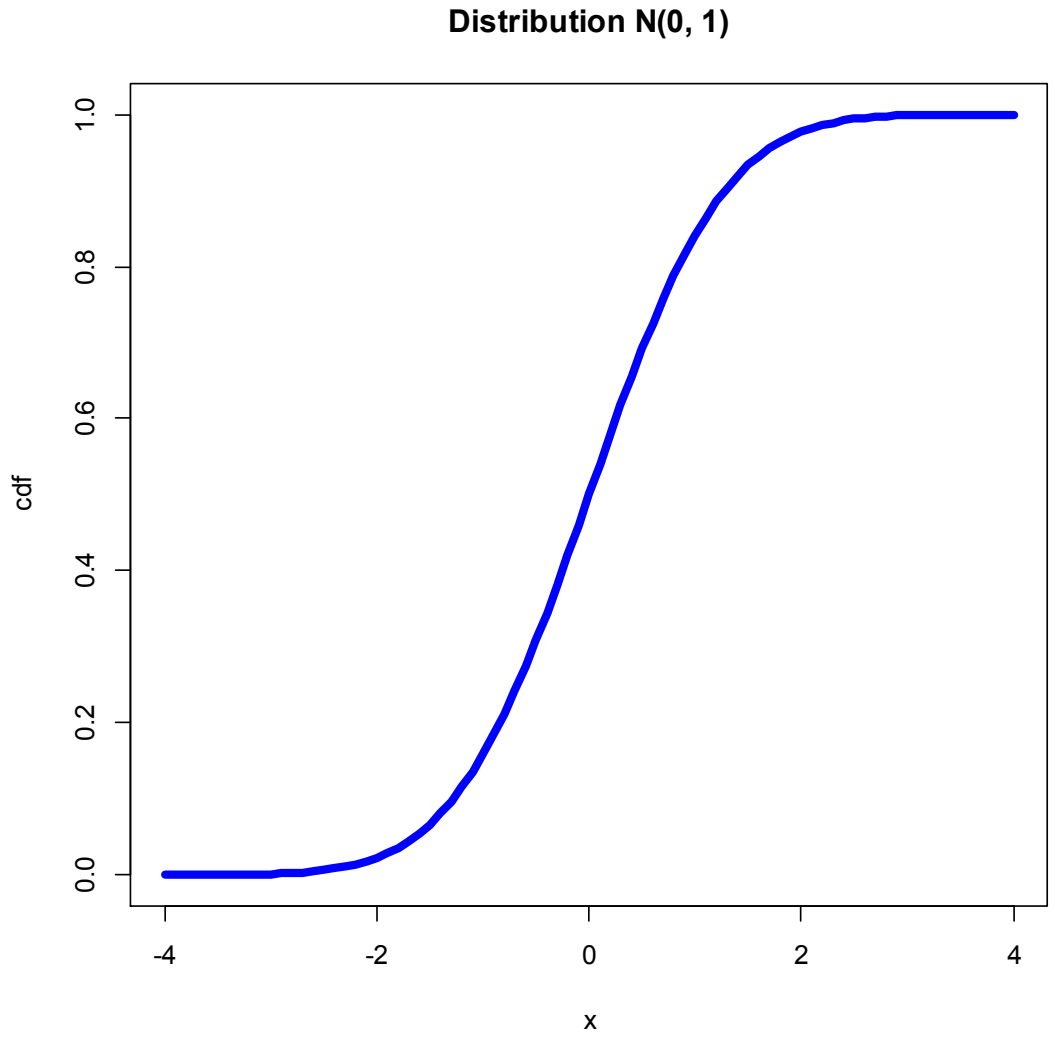
Empirical CDF with n=150 and n=170



Empirical CDF with n=190 and n=210



The Normal CDF (again)



How can we use this?

- The fact that the empirical CDF converges to the CDF of the sampling distribution, means that we can estimate moments of the sampling distribution using samples.

- The idea behind the bootstrap is to use resampling to create many "datasets" of the same size as our actual dataset, then estimate quantities of interest through these.

- The next section will discuss an instance of this idea to estimate the standard deviation of the sample mean.

Bootstrap Confidence Intervals from the Central Limit Theorem

Using the bootstrap

- The simplest way is to use the bootstrap is to calculate the variance of the statistic of interest along with the approximation given by the CLT.
- Returning to the previous example on UK Gas consumption, we may create a 95% confidence interval for the mean.

$$[\bar{X} + z_{0.025}\hat{se}_{boot}, \bar{X} + z_{0.975}\hat{se}_{boot}]$$

standard error (“deviation”) obtained by bootstrap

The (Nonparametric) Bootstrap

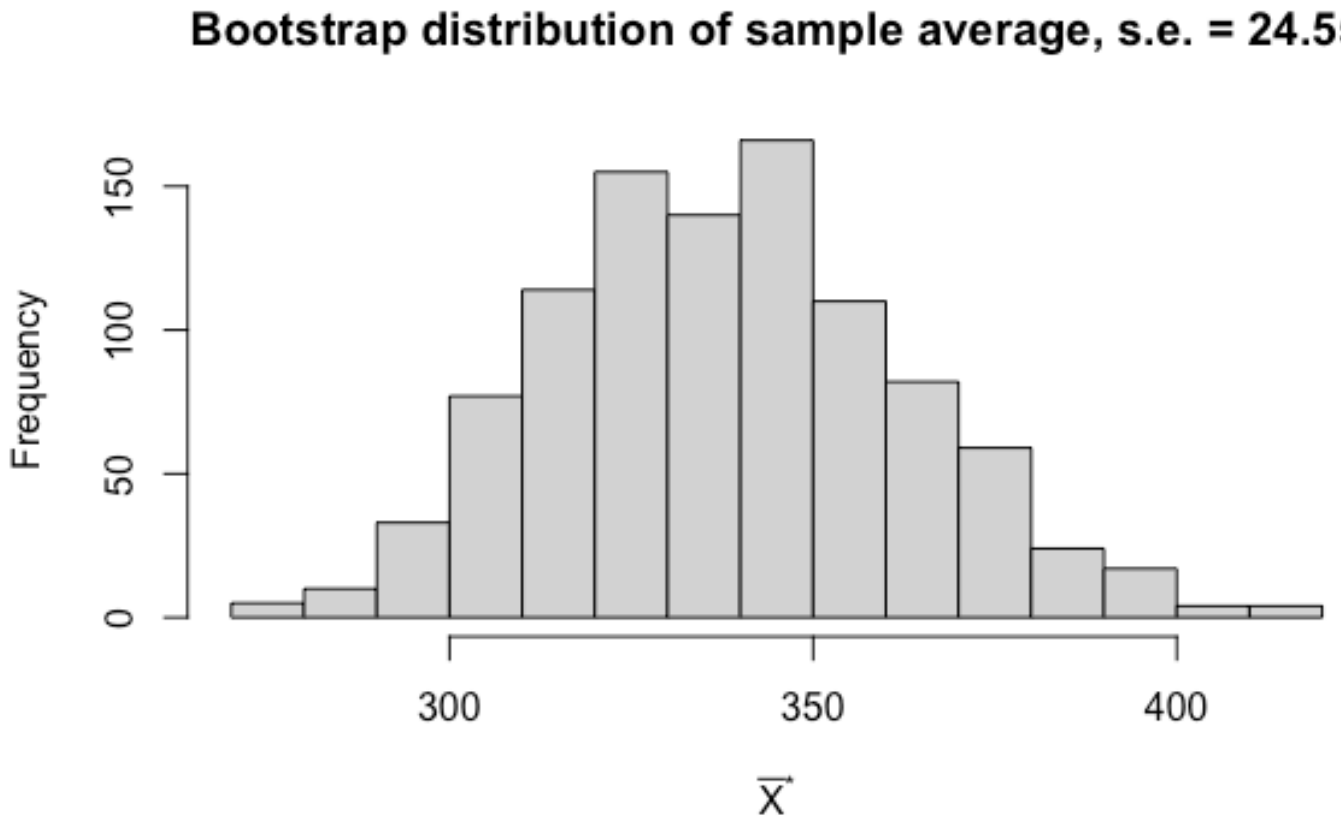
1. Draw $X_1^\star, \ldots, X_n^\star \sim \hat{F}_n$
2. Compute $\bar{X}_n^\star$ by averaging $X_1^\star, \ldots, X_n^\star$
3. Repeat steps 1 and 2, B times, to get $\bar{X}_{n,1}^\star, \ldots, \bar{X}_{n,B}^\star$
4. Let

$$s.e._{boot} \equiv \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(\bar{X}_{n,b}^\star - \frac{1}{B}\sum_{r=1}^{B}\bar{X}_{n,r}^\star\right)^2}$$

Number of bootstrap samples, B

- Ideally, we would average over *every possible re-sample*.
  - This is in general not feasible.
- So, B allows us to make an(other) approximation.
  - Again, this may be justified by the law of large numbers.
- The R demonstration includes an application of the bootstrap to a confidence interval for the mean.

UK Gas: Bootstrap Interval of the Mean



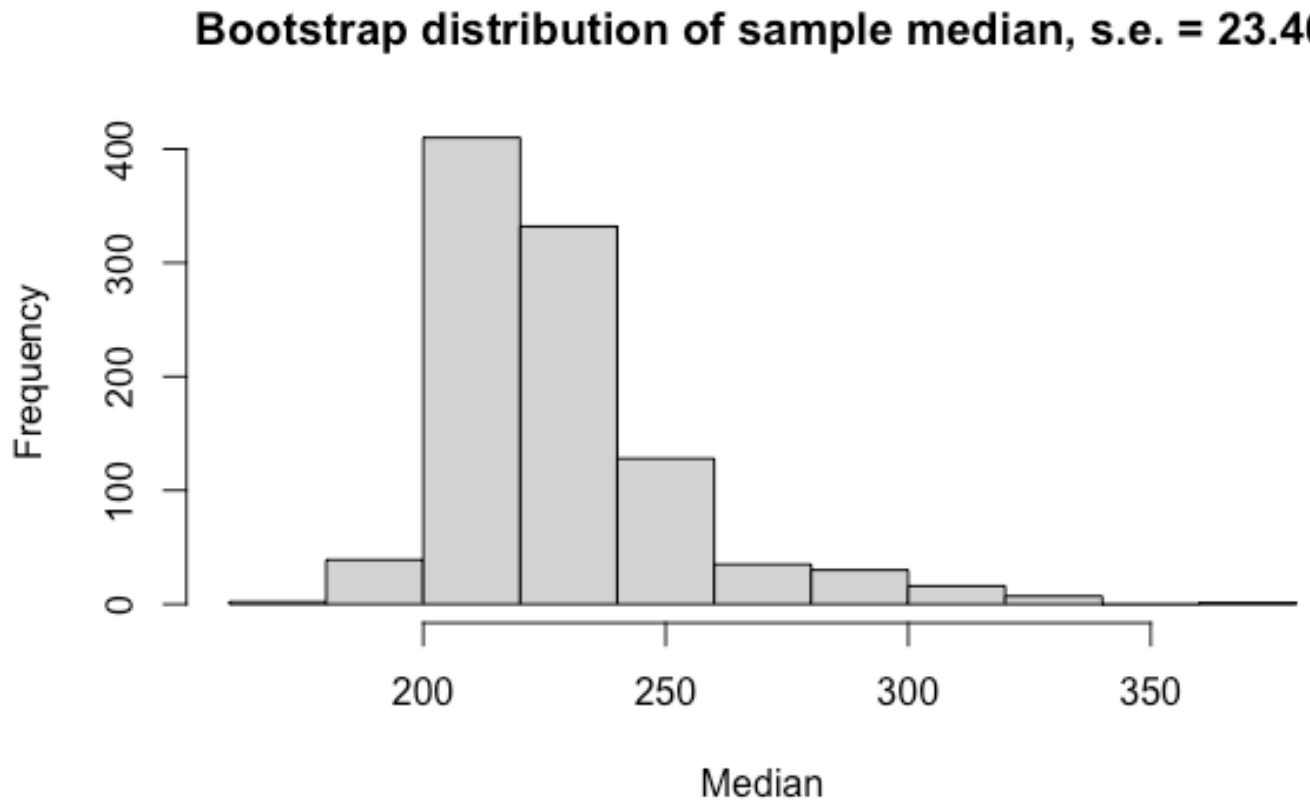

```
> cat(sprintf("0.95 Bootstrap + Normal confidence interval: [%.2f, %.2f]\n", L, U))
0.95 Bootstrap + Normal confidence interval: [289.52, 385.74]
> cat(sprintf("0.95 CLT variance + Normal confidence interval: [%.2f, %.2f]\n", L, U))
0.95 CLT variance + Normal confidence interval: [290.23, 385.03]
```

Quantities other than the mean

- The empirical mean is simple enough that we can find a decent approximation for its variance in the literature.
  - We could take the standard error of the mean to be the empirical standard deviation divided by the square root of the sample size, $\sqrt{s^2/n}$.
  - As a result, there wasn't really a need for the bootstrap in this case.

- The bootstrap shines in situations where there does not exist a theoretical approximation to the variance.
  - The UK Gas data is heavily skewed, we may therefore be more interested in the median and a confidence interval thereof.

UK Gas: Bootstrap Interval of the Median




```
> cat(sprintf("0.95 Bootstrap + Normal confidence interval: [%.2f, %.2f]\n", L, U))
0.95 Bootstrap + Normal confidence interval: [175.03, 266.77]
```

Bootstrap for a general statistic

- The idea remains as before, our statistic is now the sample median.
  - This example can be seen in the R demonstration.
  - We could apply the same technique to a general statistic T.

1. Draw $X_1^\star, \ldots, X_n^\star \sim \hat{F}_n$
2. Compute $T_n^\star$ from $X_1^\star, \ldots, X_n^\star$ according to its definition
3. Repeat steps 1 and 2, B times, to get $T_{n,1}^\star, \ldots, T_{n,B}^\star$
4. Let

$$s.e._{boot} \equiv \sqrt{\frac{1}{B}\sum_{b=1}^{B}\left(T_{n,b}^\star - \frac{1}{B}\sum_{r=1}^{B} T_{n,r}^\star\right)^2}$$

Bootstrap Pivotal Intervals

Ditching the normal assumption

- The more general bootstrap presented on the previous slide may still not be general enough.
  - Estimate the statistic.
  - Estimate the standard error of the statistic via the bootstrap.
  - Use the appropriate quantiles of the normal distribution to build a confidence interval.
- The statistic may not be (approximately) normally distributed.
  - The final step of the above approach then fails to produce a confidence interval with the correct coverage.

The bootstrap pivotal interval

- We introduced the concept of a pivot by example earlier.
  - A pivot is a function of the statistic of interest whose distribution is known.
  - There are cases in which we can't construct a pivot and the normal approximation is poor.

- One of the most used bootstrap variants of confidence intervals is the **pivotal interval**.

- Essentially, we will use the bootstrap to estimate the distribution of $\hat{\theta}_n - \theta$, which will be our pivot.
  - From the resulting empirical CDF we will determine the quantiles of interest.

The bootstrap pivotal interval

- Let H(r) be the CDF of the pivot.
  - That is, $H(r) = P(\hat{\theta} - \theta \leq r)$.

- Define quantiles such that we get coverage $1 - \alpha$:

$$P\left(a(\hat{\theta}_n) \leq \theta \leq b(\hat{\theta}_n)\right) = 1 - \alpha.$$

- You can check for yourselves that the above is satisfied by:
  - $a(\hat{\theta}_n) = \hat{\theta} - H^{-1}(1 - \alpha/2)$,
  - $b(\hat{\theta}_n) = \hat{\theta} - H^{-1}(\alpha/2)$.

The problem

- In order to determine the values of a and b we need to determine the quantiles of H, the CDF of $\hat{\theta} - \theta$.
  - This is where we invoke the bootstrap.

- Create bootstrapped pivots $R^*_{n,b} = \theta^*_{n,b} - \hat{\theta}_n$ of $\hat{\theta}_n - \theta$.
  - $\hat{\theta}_n$ is the estimate of θ using all of the data in the original sample.
  - $\theta^*_{n,b}$ is the bth bootstrap estimate of θ.

Bootstrap estimation of quantiles

- We may use the distribution of the bootstrap samples to estimate the quantile function, $\widehat{H}^{-1}(\alpha) = \theta^*_{n,B\alpha} - \hat{\theta}_n$.
  - Sort all B bootstrap pivots, $R^*_{n,b}$, in ascending order and select the one in position $B\alpha$.
  - It is likely that $B\alpha$ is not an integer, in which case simply round it to the nearest integer.

- $b(\hat{\theta}_n) = \hat{\theta}_n - \widehat{H}^{-1}\left(\frac{\alpha}{2}\right) = 2\hat{\theta}_n - \theta^*_{n,\frac{B\alpha}{2}}$
- $a(\hat{\theta}_n) = \hat{\theta}_n - \widehat{H}^{-1}\left(1-\frac{\alpha}{2}\right) = 2\hat{\theta}_n - \theta^*_{n,B\left(1-\frac{\alpha}{2}\right)}$

Take home messages

- The bootstrap pivotal interval is included within the R demonstration.

```
> cat(sprintf("0.95 Bootstrap pivot interval: [%.2f, %.2f]\n", theta_n - r_L, theta_n - r_U))
0.95 Bootstrap pivot interval: [140.80, 244.90]
```

- Estimation is important, but so is your uncertainty.
  - In applications you may be pressed for easy answers (point estimates).
  - Don’t fall for that.
- Confidence intervals provide coverage – an interval which traps the parameter of interest, regardless of its true value, with the advertised probability.

Take home messages

- However, everything is predicated on given assumptions.
  - This includes the bootstrap – you shouldn't forget this.
  - You should attempt to verify assumptions as best as you can.
  - Outside of textbook examples things are not perfect – **the aim is to be "less wrong" rather than infallible**.
- If your resulting confidence intervals appear wide and uninformative, even with large sample sizes: tough luck.
  - **Information doesn't come for free**.
  - If you want more certainty then you will need more assumptions (or more data).
- Read Chapter 9 of Wasserman for more details on the bootstrap.