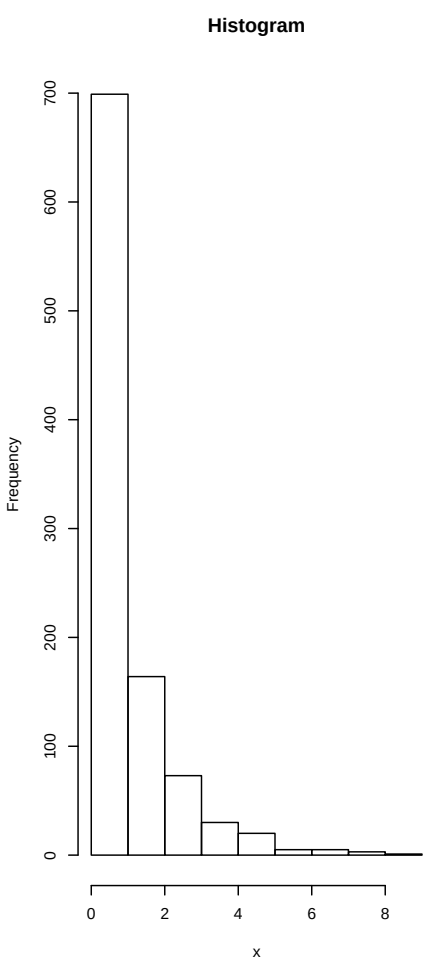
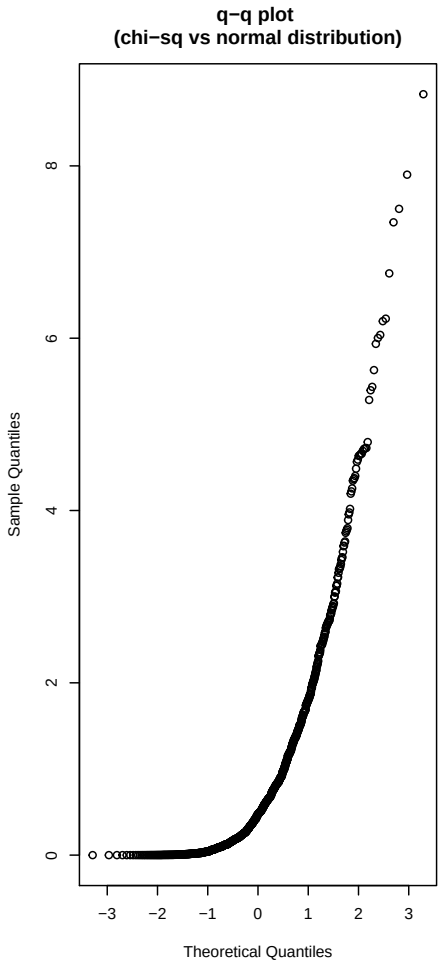
Lecture 5: Estimation and Statistical Inference

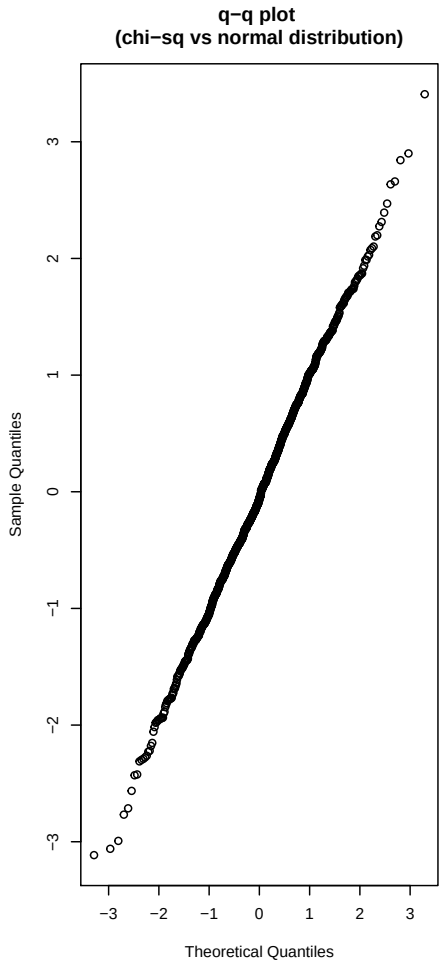
Thomas Chadefaux

PO7001: Quantitative Methods I

A final word about distributions: quantile plots





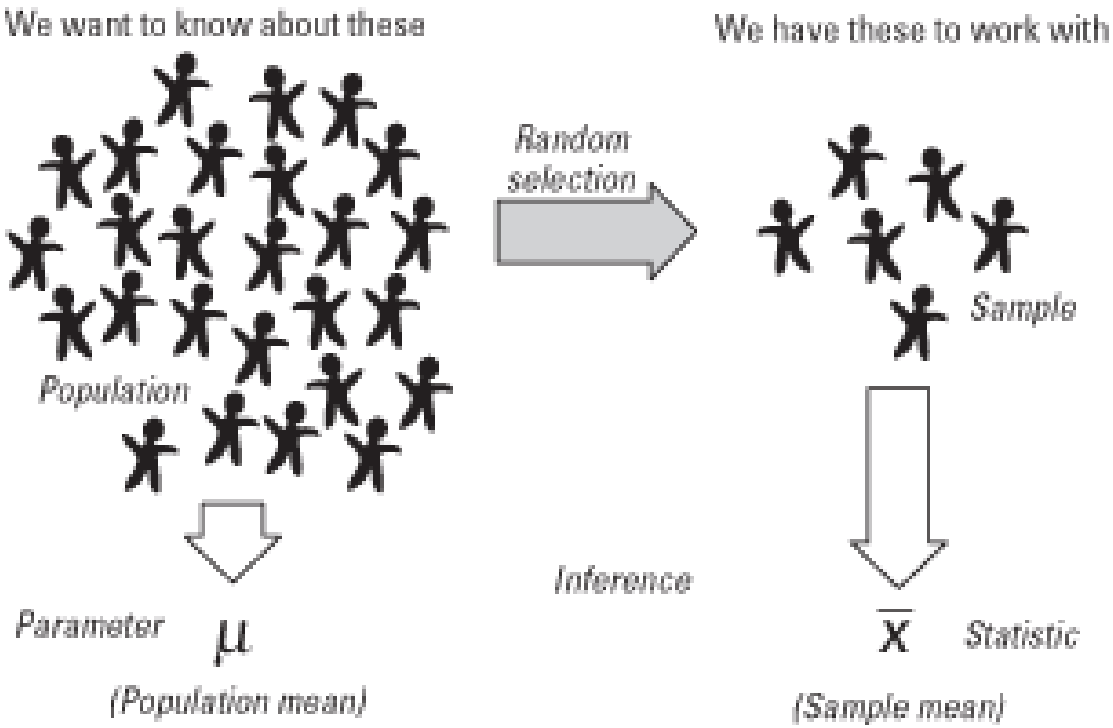


Outline of weeks 5 & 6

- What is statistical inference?
- Confidence intervals
- tests of significance
- Bootstrapping

Inference

What is Inference?



What is Inference?

- Purpose of inference: Draw conclusions from the data.
- Using our sample, we want to draw conclusions about the general population
- This involves making probability calculations to take into account chance.

An example

- Test whether a drug is effective
- Sample of 40 patients.
  - 20 given the drug
  - 20 given the placebo
- 60% of the treatment group get better (i.e., 12 of them), whereas only 8 of the placebo group get better.
- Can we conclude that the drug is effective?
- In fact, these results would occur by chance one out of every five times by sheer chance. So we conclude that the observed difference is too likely to be due to chance, and hence fail to reject the hypothesis that the drug is ineffective.

Statistical Inference

- Two main types:
  - Confidence Intervals: estimate the value of a parameter
  - Tests of significance: assess the evidence for a claim
- Both rely on assessing what would happen if we sampled many times. In particular, would we be fairly likely to obtain this result by chance?
- Because inference relies on samples, they will be based on the *sampling distribution* of the data. E.g., if we sampled over and over, how would, say, the means of those samples be distributed? So we need to know about the *sampling distribution* of the sample means.

The sampling distribution of the sample mean

- The sample mean $\bar{x}$ from a sample or experiment is an estimate of the mean μ of the underlying population (remember that greek is for population, latin for sample).
- Suppose that we collect many samples and calculate each sample's mean. We now have a set of means $\bar{x}_1, \bar{x}_2, \ldots \bar{x}_N$.
- What is the mean of those means? Note that the expected value of $\bar{x}_i$ is simply μ. So the mean of those means is simple

$$\frac{1}{N}(\bar{x}_1 + \bar{x}_2 + \ldots + \bar{x}_N)$$

$$= \frac{1}{N}(\mu + \mu + \ldots + \mu) = \mu$$

- A little more complicated, however, is the variance. I.e., what is

$$Var(\bar{X})$$

The variance of the means

- First note that $Var(aX) = a^2 Var(X)$. Why? in-class exercice: prove it!
- Second, note that variance of the sum (or the difference) of uncorrelated random variables is the sum of their variances (proof is beyond this class):

$$Var(\sum_{i=1}^{N} X_i) = \sum_{i=1}^{N} Var(X_i)$$

- Therefore, the variance of the means is:

$$Var(\bar{X}) = Var(\frac{1}{n}\sum_{i=1}^{n} X_i) = \frac{1}{n^2}\sum_{i=1}^{n} Var(X_i)$$

$$= \frac{1}{n^2}\sum_{i=1}^{n}(X_i - \bar{X})^2 = \frac{1}{n^2}\sum_{i=1}^{n}\sigma^2 = \frac{1}{n^2}n\sigma^2 = \frac{\sigma^2}{n}$$

So $\sigma_{\bar{x}}^2 = \frac{\sigma^2}{n}$ and $\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$ When $\sigma_{\bar{x}}$ is estimated from the data, we call it the *standard error SE*

(Proof that $Var(aX) = a^2 Var(X)$)

$$Var(aX) = \frac{1}{N}\sum_i (aX - a\bar{X})^2$$

$$= \frac{1}{N}\sum_i (a(X - \bar{X}))^2$$

$$= \frac{1}{N}\sum_i a^2(X - \bar{X})^2$$

$$= Na^2\frac{1}{N}\sum_i (X - \bar{X})^2$$

$$= a^2\sum_i (X - \bar{X})^2 = a^2 Var(X)$$

Statistical confidence

- Suppose for now that we know that $\sigma = 2$ (in practice we almost never do, but let us assume for now). I.e., the standard deviation in the population is 2.
- Suppose we are taking samples of size 100
- From this, we can calculate the standard deviation *of the sampling distribution* as:

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{N}} = \frac{2}{10} = \frac{1}{5}$$

- Note that we know that for a normal distribution, 95% of the data is within $\pm 2\sigma$ of the mean (more precisely, $\pm 1.96\sigma$)
- So 95% of all samples will be between $\mu - 2\sigma_{\bar{x}}$ and $\mu - 2\sigma_{\bar{x}}$,
- Suppose that we collect a sample and that sample has mean 10. So we can say that we are 95% confident that the unknown *population* mean is between $\mu - 2\sigma_{\bar{x}}$ and $\mu + 2\sigma_{\bar{x}}$ i.e., between $10 - 2/5 = 9.6$ and $10 + 2/5 = 10.4$.

Confidence Intervals

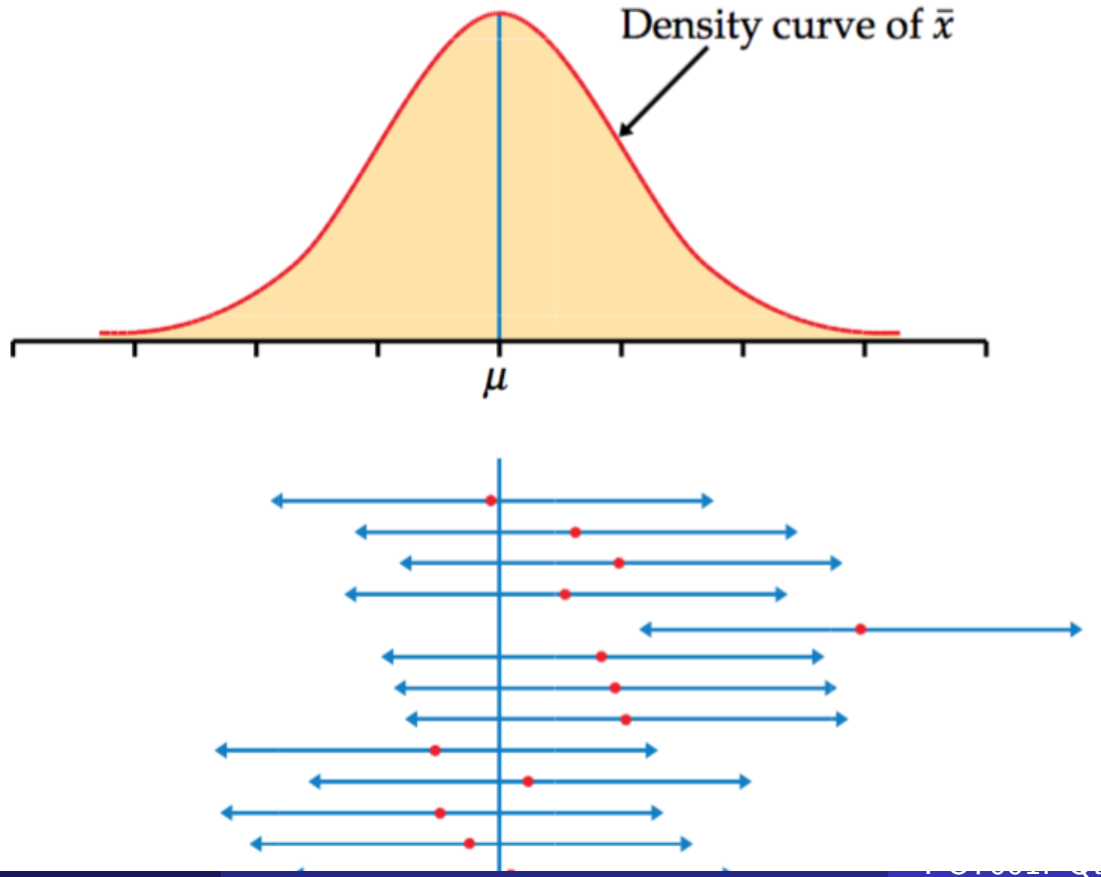
What we just calculated is a 95% confidence interval for the mean μ. Like most CIs, it has the form:

$$\text{estimate} \pm \text{margin of error}$$

The estimate, in this case, was the mean. The margin of error reflects how accurate we think our estimate is.

Interpretation

The 95% confidence interval tells us that if we sampled over and over, and every time calculated the 95% confidence interval of the mean based on that sample, then 95% of these intervals would contain the true value of μ



Calculating the CI

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$$

- For example: Suppose the sample mean is 20 and the standard deviation is 10. We have 900 observations. Then the confidence interval of the mean is:

$$20 \pm 1.96 \times \frac{10}{30} \approx [19.33, 20.66]$$

- Note that the size of the confidence interval depends critically on:
  - the significance level of the test (here 5%), and, correspondingly, the critical value of the test (1.96 here)
  - the sample size n. To see this, consider the example above, but this time with only 25 observations:

$$20 \pm 1.96 \times \frac{10}{5} \approx [16, 24]$$

I.e., a much larger interval