PO 7005
Lecture 1. Statistics: a Review

Thomas Chadefaux

Contents

1 Introduction to the Course

1.1 Overview

One big theme: we have a population—countries, people, districts, voters, etc., and we want to learn something about that population. The problem, however, is that we usually cannot observe the entire population. Instead, we have a sample—a subset—of that population. Based on this sample, we need a technique to make inferences about the entire population.

For example, we'll want to estimate the following model:

$$\text{Political orientation}_i = \alpha + \beta_1 \text{education}_i + \beta_2 \text{Wealth}_i + \varepsilon_i$$

and get estimates of parameters α, β_1 and β_2. In general, these sample estimates will differ from the population true parameters because our sample is usually not perfectly representative of the entire population. I.e., we have sampling error. Our goal will be to find techniques that reduce this sampling error. I.e., we want to get estimates b_1 that are as close as possible to β_1.

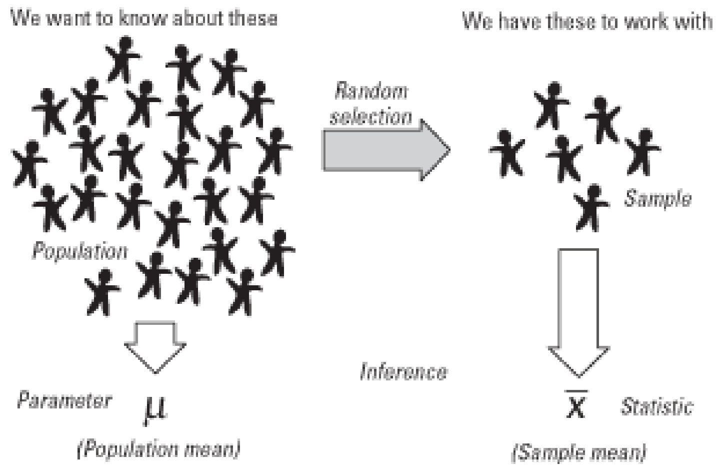


Figure 1: Population and Sample

The first technique we look at is Ordinary Least Squares (OLS). Why OLS? Because given a number of assumptions (the Gauss-Markov assumptions), it has many desirable properties. I.e., it is a very good tool (it is 'BLUE'—more on this later) to make inferences about the population based on the sample.

But of course, we want to make sure that the assumptions we made are correct. We'll need diagnostic tests. If the assumptions are not satisfied, then OLS is no longer 'ideal', and so we'll need another tool. That's why we'll then talk about other techniques such as instrumental variables, generalised least squares, maximum likelihood, etc.

Finally, we'll be interested in looking at data over time. Panel data and time series will conclude this class.

1.2 Econometrics vs. 'Hard Science'

In hard sciences, we might want to know whether some fertiliser contributes to plant growth. We have an experiment pot and a control pot. By comparing the two, we then know whether fertiliser causes plant growth.

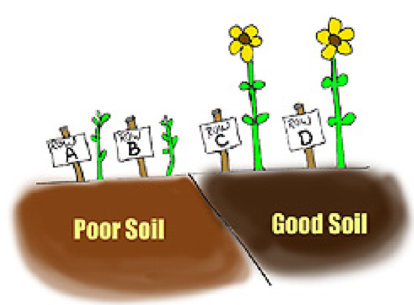


Figure 2: An Experiment

In social sciences, however, we usually cannot conduct these experiments. There are several reasons for this.

- Ethical reasons. Studying war...
- Practical reasons. How do you 'assign' a family? A socio-economic background?
- Endogeneity, reverse causality, missing variables, selection bias, etc. For example, we may find that people who listen to NPR vote more to the left than those who don't. But the causation probably runs the other way, or at least both ways. Or we might find that countries that trade a lot tend not to fight. But these countries also tend to share alliances and to be democracies.

The importance of natural experiments: Ideally we'd like some form of natural experiment, so look out for them. Two examples:

- Angrist 1990: interested in the effect of participation in war on lifetime income. Problem is that participation in war is often not random. I.e., only a certain type of people enroll in the military, and their type might affect their earning ability in the first place. The Vietnam war, however, provided a natural experiment: draft was based on birth day (which day of year). For example, every young male born on July 23 would be drafted. Since birthdate is unrelated to income (actually this is debatable—see Gladwell's argument), it was a way to control for selection bias.
- Another experiment addresses the problem of feasibility.[1] For example, we might be interested in how to alleviate poverty, and wonder how to go about it. Is it better to give a small amount to many people, or a very large amount to a few people. Problem is that it would be very costly to conduct such an experiment, and even then we would have to wait years/decades to see the effect. Bleakley and Ferrie analyse the effect of the Georgia's Cherokee Land Lottery of 1832.

[1] "Shocking Behavior: Random Wealth in Antebellum Georgia and Human Capital Across Generations". `http://www.economics.illinois.edu/seminars/development/documents/Bleakley_Paper2.pdf`. Here is the abstract:

> "Does the lack of wealth constrain parents' investments in the human capital of their descendants? We conduct a fifty-year followup of an episode in which such constraints would have been plausibly relaxed by a random allocation of wealth to families. We track descendants of those eligible to win in Georgia's Cherokee Land Lottery of 1832, which had nearly universal participation among adult white males. Winners received close to the median level of wealth—a large financial windfall orthogonal to parents' underlying characteristics that might have also affected their children's human capital. Although winners had slightly more children than non-winners, they did not send them to school more. Sons of winners have no better adult outcomes (wealth, income, literacy) than the sons of non-winners, and winners' grandchildren do not have higher literacy or school attendance than non-winners' grandchildren. This suggests only a limited role for family financial resources in the formation of human capital in the next generations in this environment and a potentially more important role for other factors that persist through family lines."

Today's class is a refresher on some of the fundamental concepts needed for the rest of the term.

2 Random Variables

- Random variable: a variable whose value is determined by chance.[2,3]
- A probability density function (PDF) assigns a probability to each value of a random variable X.[4]
  - A discrete PDF for a variable X taking on the values $x_1, x_2, \ldots, x_n$ is a function such that:

$$f_X(x) = P[X = x] \text{ for } i = 1, 2, 3, \ldots, n$$

    and 0 otherwise.[5]
  - Similarly, a continuous PDF is a function such that:[6]

$$P[a < x < b] = \int_a^b f(x)dx$$

    (see normalPDF.R)

    Test yourself: can a PDF take on a value greater than 0?
- The cumulative probability density function (CDF) gives the probability of a random variable being less than or equal to some value:
  - Discrete case: $F(x) = \sum_{x_j < x} f(x_j)$
  - Continuous case: $F(x) = P[X < x] = \int_{-\infty}^{x} f(u)du$

In-class exercise: Replicate the plots on the right in R

[2] A little more formally, a random variable is a function $X : \Omega \to E$, where Ω denotes the set of all possibly outcomes and E is a set. For example, suppose that X represents the random variable 'flip a coin twice'. Then $\Omega = \{\{H, H\}, \{H, T\}, \{T, H\}, \{TT\}\}$ and $\{H, T\}$, $\{T, T\}$, etc., are realisations of the random variable X.

[3] Convention: a random variable is denoted by an upper case letter (e.g., X), and realisation of that random variable by lower case (e.g., $x_1, x_2, \ldots, x_n$)

[4] Technically we should be talking about a probability *mass* function for discrete variables, and a probability *density* function for the continuous case.

[5] Why do we bother writing $f_X(x)$ and not just $f(X)$? It's a matter of convention, but the idea is that we want to remember that $f_X(x) = P[X = x_i]$. Without it, we might be confused when we evaluate, say, $f(2)$, which makes no sense. $f_X(2)$, on the other hand, means that we are looking for the probability that X is 2 (i.e., $P[X = 2]$).

[6] In the continuous case, the mass of $P(X = x_i)$ is 0. We need to integrate over an interval to get a probability.

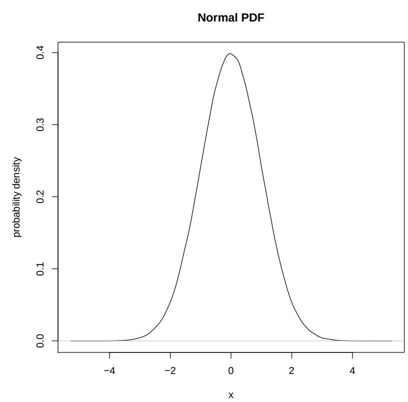


Figure 3: normalPDF.R

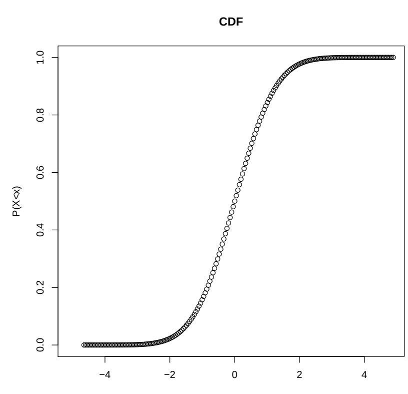


Figure 4: normalCDF.R

3 Joint Probability Density Functions

- Joint probability density function:

$$f(x, y) = P(X = x \text{ and } Y = y]$$

- The marginal PDF is a function that returns the probability to observe x, without consideration of the value of y :

$$f_X(x) = P(X = x) = \sum_y f_{X,Y}(x, y)$$

- The conditional PDF gives us the probability that $X = x$ given that $Y = y$:

$$f_{X,Y}(x|y) = P(X = x|Y = y) = \frac{f_{X,Y}(x, y)}{f_Y(y)}$$

- X and Y are **statistically independent** iff:

$$f_{X,Y}(x, y) = f_X(x) \cdot f_Y(y)$$

  In more intuitive terms, this means (loosely) that one event conveys no information about the other. The probability of event x occurring is independent of the occurrence of event y.

  In-class exercise: Generate:

  - two variables that are statistically independent
  - two variables that are statistically dependent.
  - two uncorrelated but statistically dependent variables.

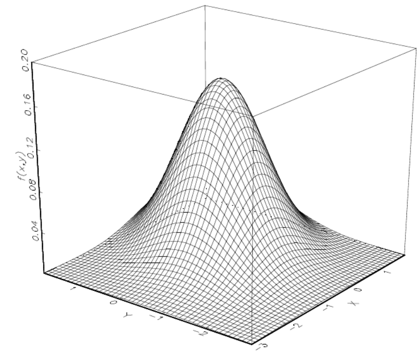


Figure 5: Joint PDF

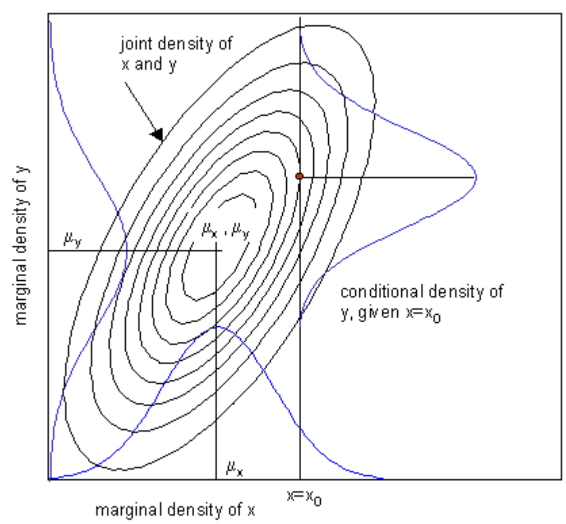


Figure 6: Putting it all together

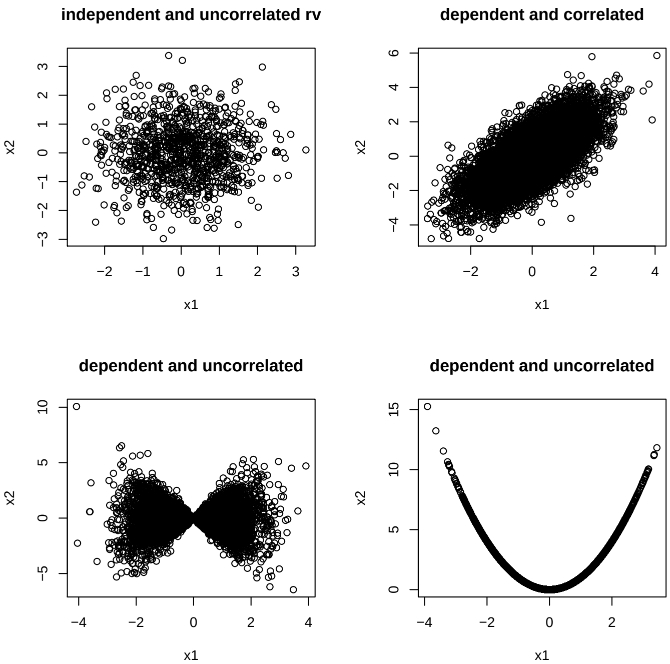


Figure 7: statisticalIndependence.R

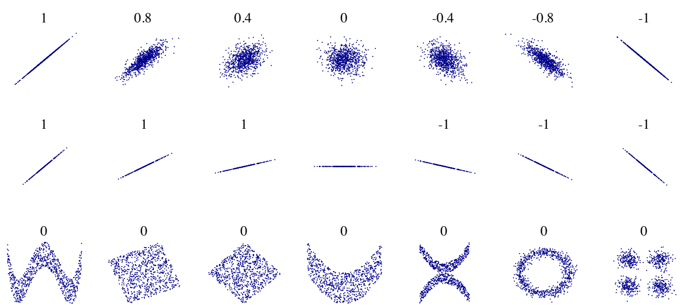


Figure 8: Note that variables that are uncorrelated are not necessarily independent. From Wikipedia: Several sets of (x, y) points, with the Pearson correlation coefficient of x and y for each set. Note that the correlation reflects the noisiness and direction of a linear relationship (top row), but not the slope of that relationship (middle), nor many aspects of nonlinear relationships (bottom). N.B.: the figure in the center has a slope of 0 but in that case the correlation coefficient is undefined because the variance of Y is zero.

4 Expectations, Variance and Covariance

4.1 Expectations

The expected value is the average value that a random variable takes on over many repeated trials. I.e.,

If all events are equally likely, then $E[X] = \frac{1}{n}\sum_{i=1}^{n} x_i$

$$E[X] = \sum_{i=1}^{n} x_i f_X(x) \quad \text{if } X \text{ is discrete}$$

$$E[X] = \int_{-\infty}^{\infty} x f_X(x) dx \quad \text{if } X \text{ is continuous}$$

For example, the expected value of random variable "throw a die and record the outcome" is:

$$1 \times \frac{1}{6} + 2 \times \frac{1}{6} + \ldots + 6 \times \frac{1}{6} = 3.5$$

Note that the expected value is the long-run average of repeated experiment. I.e., suppose we threw the die an infinity of times, then the average would equal the expected value.

- The expected value of a constant is itself: $E[b] = b$
- Useful to know for future derivations: If a and b are constants, then $E[aX + b] = aE[X] + b$

4.2 Variance and covariance

The expected value of a random variable gives a crude measure of the central measure of that variable. But we also want a measure of spread. The variance is one such measure.

- The *variance* measures the distribution of values of X around its expected value $E[X]$:

  $$var(X) = \sigma_x^2 = E[(X - E[X])^2] = \sum_x (X - E[X])^2 f(x)$$

  The standard deviation is the square root of the variance:

  $$\sigma_X = \sqrt{var(X)}$$

- The *covariance* of X and Y is :

  $$cov(X, Y) = E[(X - E[X])(Y - E(Y)]$$

The population variance can also be written as

$$\sigma_x^2 = E[X^2] - \mu^2$$

Proof:

$$\begin{aligned}\sigma_X^2 &= E[(X - E[X])^2] \\ &= E[(X^2 - 2\mu X + \mu^2)] \\ &= E[X^2] - 2\mu E[X] + \mu^2 \\ &= E[X^2] - 2\mu\mu + \mu^2 \\ &= E[X^2] - \mu^2\end{aligned}$$

Note that the variance is just the covariance of X with itself:

$$\begin{aligned}var(X) &= E[(X - E[X])^2] \\ &= E[(X - E[X])(X - E[X])] \\ &= cov(X, X)\end{aligned}$$

Note that this can be rewritten as:

$$\begin{aligned} cov(X,Y) &= E[(X-E[X])(Y-E[Y]] \\ &= E\left[XY - XE[Y] - E[X]Y + E[X]E[Y]\right] \\ &= E[XY] - E[X]E[Y] - E[X]E[Y] + E[X]E[Y] \\ &= E[XY] - E[X]E[Y] \end{aligned}$$

Note that if X and Y are independent, then $E[XY] = E[X]E[Y]$, and so

$$cov(X,Y) = E[XY] - E[X]E[Y] = 0$$

- Problem: The size of the covariance depends on the units in which X and Y are measured (see e.g. on the right). This has led to the use of the correlation coefficient $\rho \in [-1,1]$:
$$\rho = \frac{cov(X,Y)}{\sigma_x \sigma_y}$$
- Variance of correlated variables (why? see proof on the right):
$$var(X+Y) = var(X) + var(Y) + 2cov(X,Y)$$
$$var(X-Y) = var(X) + var(Y) - 2cov(X,Y)$$

e.g.,

```
n <- 100000

x <- rnorm(n)
y <- x +rnorm(n)
cov(x,y) # returns ~1

x <- x *2
y <- y *2
cov(x,y) # returns ~4
```

Proof:

$$\begin{aligned} & Var(X+Y) = && (1)\\ &= E(X+Y-E(X+Y))^2 && (2)\\ &= E((X+Y)^2 - 2(X+Y)E(X+Y) + (E((X+Y))^2) && (3)\\ &= E((X+Y)^2) - (E(X+Y))^2 && (4)\\ &= E(X^2) + 2E(XY) + E(Y^2) - (E(X+Y))^2 && (5)\\ &= E(X^2) + 2E(XY) + E(Y^2) - (E(X)+E(Y))^2 && (6)\\ &= E(X^2) + 2E(XY) + E(Y^2) - (E(X)^2 + 2E(X)E(Y) + E(Y)^2 && (7)\\ &= (E(X^2) - E(X))^2 + (E(Y^2) - (E(Y))^2 + 2E(XY) - 2E(X)E(Y) && (8)\\ &= Var(X) + Var(Y) + 2Cov(X,Y) && (9)\\ & && (10) \end{aligned}$$

4.3 A note on correlation, linear independence and orthogonality

Two vectors are:

- Uncorrelated iff: $(\mathbf{X} - \bar{X})'(\mathbf{Y} - \bar{Y})' = 0$
- Linearly independent iff there is no constant a such that $a\mathbf{X} - Y = 0$.
- Orthogonal iff $\mathbf{X}'\mathbf{Y} = 0$.

First consider linear independence. In R^2, linear independence implies that one vector is a (linear) function of the other. Consider for example the vectors $v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $v_2 = \begin{pmatrix} 2 \\ 4 \end{pmatrix}$. Then $v_2 = 2v_1$ and so $\exists a \neq 0$ s.t. $av1 - v2 = 0$ (in this case, $a = 2$). So these two vectors are not linearly independent.

Contrast this with the case in which the two vectors are linearly independent, that is, $v1 \neq av2$ for all $a \neq 0$. For example, $v_1 = \begin{pmatrix} 1 \\ 2 \end{pmatrix}$, $v_2 = \begin{pmatrix} 2 \\ 5 \end{pmatrix}$.

ON CORRELATION, independence and orthogonality: A clear explanation is in Rodgers, Nicewander and Toothhaker (1984). Linearly Independent, Orthogonal, and Uncorrelated Variables. The American Statistician 38(2):133–134:

> "Each variable is a vector lying in the observation space of n dimensions. Linearly independent variables are those with vectors that do not fall along the same line; that is, there is no multiplicative constant that will expand, contract, or reflect one vector onto the other. Orthogonal variables are a special case of linearly independent variables. Not only do their vectors not fall along the same line, but they also fall perfectly at right angles to one another (or, equivalently, the cosine of the angle between them is zero). The relationship between "linear independence" and "orthogonality" is thus straightforward and simple.
>
> Uncorrelated variables are a bit more complex. To say variables are uncorrelated indicates nothing about the raw variables themselves. Rather, "uncorrelated" implies that once each variable is centered (i.e., the mean of each vector is subtracted from the elements of that vector), then the vectors are perpendicular. The key to appreciating this distinction is recognizing that centering each variable can and often will change the angle between the two vectors. Thus, orthogonal denotes that the raw variables are perpendicular. Uncorrelated denotes that the centered variables are perpendicular."

Consider now correlation, in particular Pearson's correlation coef: $\rho = \frac{(\mathbf{v_1}-\bar{\mathbf{v}}_\mathbf{1})'(\mathbf{v_2}-\bar{\mathbf{v}}_\mathbf{2})}{\sigma_{v_1}\sigma_{v_2}}$ Note that this is 0 for orthogonal vectors, 1 or -1 for linearly dependent vectors, and anywhere between 0 and 1 for linearly independent vectors.

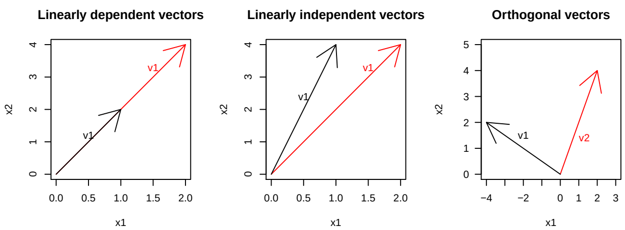


Figure 9: IndependenceEtc.R

5 A refresher on inference

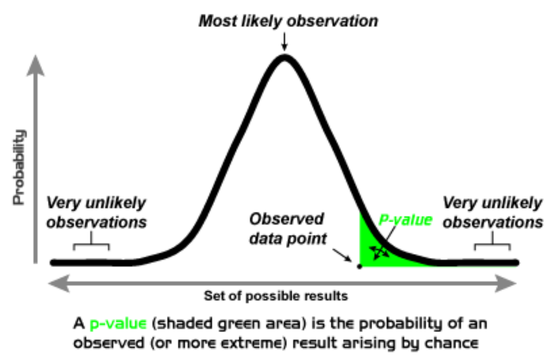
Note: this is a very broad refresher. It does not replace proper understanding of each of the concepts and methods from Quants 1!

Suppose we collected a sample of IQ scores from students and want to compare it to the IQ scores of the overall population. We calculate the average IQ scores of our students (say, 105), and would like to know whether this point estimator is "compatible" with the hypothesized version of the average IQ score (here 100). The null hypothesis will be that $IQ_{students} = IQ_{pop}$, and the alternative hypothesis may be $IQ_{students} \neq IQ_{pop}$.

We know that if we sampled over and over, the sample means would have a certain distribution with mean 100 and some standard deviation, which we call the standard error.. To estimate that standard deviation of the sample means (i.e., the standard ERROR), we can use the formula (see quants 1)

the standard error is the standard deviation of the sample statistics. This is NOT to be confused with the SD of the sample!

$$SE = \frac{s}{\sqrt{n}}$$



So with this, we know the probability distribution of the sample means under the Null hypothesis, i.e., we know that $\bar{X} \sim N(100, SE)$. (Question: why is it normally distributed?). Now the only question left is whether our observed sample mean is 'extreme' (i.e., unlikely) given what we know about the distribution of sample means. To determine this, we calculate a Z value, which is really just a normalized version of our mean:

$$Z = \frac{\bar{X} - \mu}{SE}$$

. Then we calculate the probability to observe such a large (or small) Z *if* the null hypothesis were true, and on that basis either reject or fail to reject the null hypothesis.

In Class: Example

6 Important Distributions

Different statistical tests rely on different distributions, depending on how the underlying statistic is distributed. For the mean we used the normal distribution (well, technically the t but anyway), but there will be cases where we need other distributions. This is just a brief overview

6.1 The Normal distribution

A random variable X is normally distributed ($X \sim N(\mu, \sigma^2)$) if it has the PDF

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e\left(-\frac{x-\mu^2}{2\sigma^2}\right),$$

where μ is the mean and σ^2 the variance.

6.2 χ^2: The Chi-squared distribution

Let $Z_1, Z_2, \ldots, Z_n$ be independent standard normal distributions. Then

$$Y = \sum_{i=1}^{n} Z_i^2$$

has a chi-squared distribution with n degrees of freedom (i.e., $Y \sim \chi_n^2$).

Later, we will talk about the sum of squared residuals. Since residuals are assumed to be normally distributed, their square should be distributed χ.

In-class exercise: Replicate the plot on the right

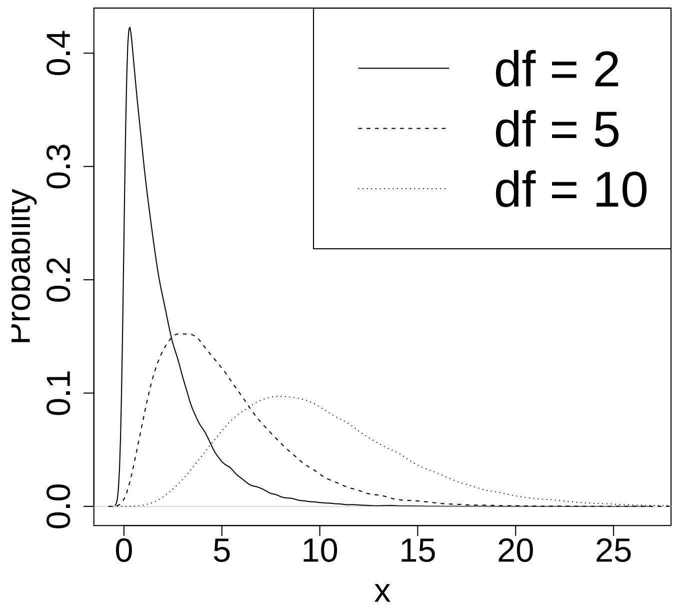


Figure 10: χ^2 distribution. See chiSquared.R

6.3 Student's t distribution

Suppose $Z \sim N(0,1), U \sim \chi_n^2$, and U and Z are independently distributed. Then the variable

$$t = \frac{Z}{\sqrt{U/k}}$$

has a t distribution with k degrees of freedom.

Why should we care about the t distribution? We are often interested in the probability of observing a given sample mean. For example, a sample of people have received a treatment and we want to know whether that treatment is effective. For example, are these people more likely to donate to a political party if they have watched a campaign, i.e., $\bar{d} > 0$? To determine whether this is the case, we need

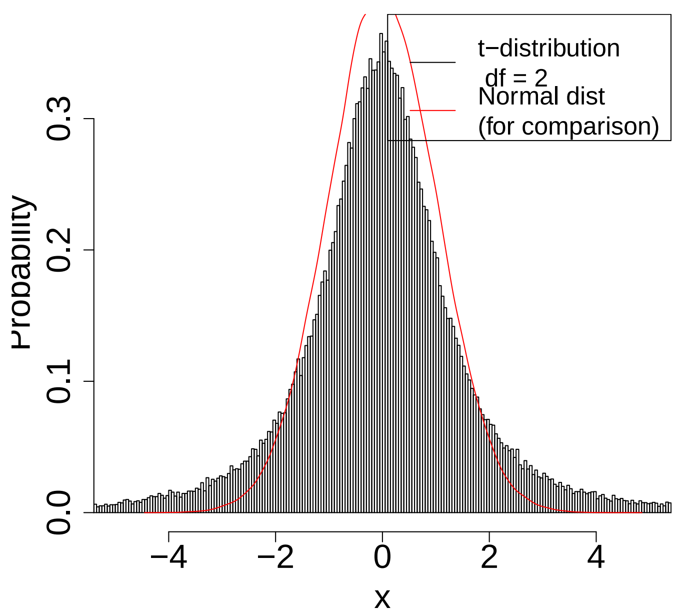


Figure 11: t distribution. See tDistribution.R

to compare $\bar{d}$ to the null hypothesis, i.e., $\bar{d} = 0$, while taking into account the standard deviation of the *sampling distribution*. To do that, we need to know how many standard deviations above the mean our sample mean is. To do this, we calculate

$$\frac{\bar{d} - 0}{\sigma_d}$$

, where σ_d is the standard deviation of the sampling distribution (i.e., the standard error). The problem is that we usually do not know σ_d. However, we can estimate this standard error of the mean by

$$\sigma_d = \frac{s}{\sqrt{n}}$$

Why $\sigma_x = \frac{s}{\sqrt{n}}$? Suppose we have n observations from a variable $X \sim N(\mu, \sigma^2)]$. Let $T = (X_1 + X_2 + \ldots + X_n)$. Then $\sigma_T^2 = n\sigma^2$, and $\sigma_{T/n}^2 = \frac{1}{n^2} n\sigma^2 = \frac{\sigma^2}{n}$. So the standard deviation of T/n is $\sigma_{T/n} = \frac{1}{n^2} n\sigma^2 = \frac{\sigma}{\sqrt{n}}$. Now note that T/n is the sample mean, and so we just calculated the standard deviation of the sample mean, i.e., the standard error of the mean. See also Dougherty p. 25.

SO, to know how many standard deviations above the mean our sample mean is, we calculate

$$\frac{\bar{d} - 0}{\sigma_d} = \frac{\bar{d}}{(\sqrt{s^2/n}}),$$

which follows a t-distribution (why?).

6.4 The F distribution

The F distribution is the distribution of the ratio of two independent chi-squared random variables divided by their respective degrees of freedom. For example, let $U \sim \chi_m^2$ and $V \sim \chi_N^2$, then the variable

$$F = \frac{U/m}{V/n}$$

has an F distribution and n degrees of freedom. I.e., $F \sim F_{m,n}$.

This will be useful for the F-test in the context of regression. The nominator and the denominator will be residual sums of squares, i.e., chi-square distributed variables.

Why do we care? Often we will want to compare two regression models, where model 1 is 'nested' within model 2. We want to know if model 2 is better than model 1 (i.e., the additional variable(s) was worth it). Model 1 has p_1 parameters, whereas model 2 has p_2 parameters. U and V will be the sum of squared residuals from each regression (i.e., two chi-squared variables). The ratio therefore follows an F distribution.

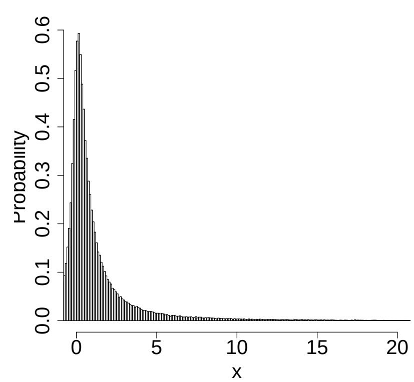


Figure 12: F distribution. See Fdistribution.R

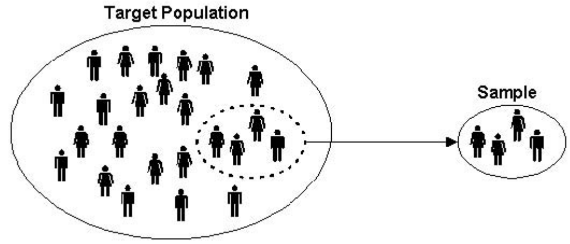
7 Estimators

Question: how can we infer the population parameters (e.g., mean, σ) from a sample?

We need a method that uses the data from the sample to estimate the population parameters. We use estimators, which can be understood as functions that take data from our sample, and return an estimate. I.e., *estimator* : *sample* $\rightarrow \beta^*$

Because of sampling error, our estimate will not be exactly equal to the population parameter. But suppose we sampled over and over from the population and for each of these samples calculated a number of statistics. I.e., what is a "good" estimator? Some of the attributes we'd like to have are:

- Unbiasedness: $E[\beta^*] = \beta^P$, where β^P is the true population parameter.
- Consistency: the larger my sample size, the closer I get to the true value. Formally, an estimator of parameter θ is consistent if:$\text{plim}_{n\rightarrow\infty}\tilde{\theta}_n = \theta$. Note that this is the large sample analog of unbiasedness.
- Efficiency: we want an estimator with low variance. I.e., we prefer β^* to $\tilde{\beta}$ if $var(\beta^*) < var(\tilde{\beta})$.



An estimator can be unbiased but not consistent. For example, we can estimate the mean of a sample $\{x_1, x_2, \ldots, x_n\}$ by $\tilde{\mu} = x_1$. This estimator is unbiased, as $E[x_1] = \mu$, but not consistent, as it does not converge to any value. However, if a sequence of estimators is unbiased and converges to a value, then it is consistent, as it must converge to the correct value.

An estimator can be biased but consistent. For example, if we estimate the population mean by $\tilde{\mu} = \frac{1}{n}\sum x_i + \frac{1}{n}$, our estimator is biased (since $E[\tilde{\mu}] = \mu + \frac{1}{n}$), but consistent because as n increases, $\frac{1}{n}$ becomes smaller and smaller.

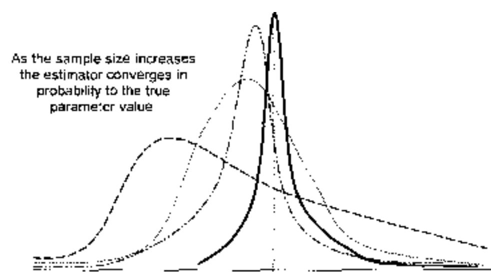


Figure 13: Consistency

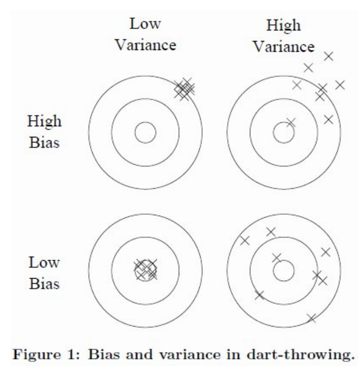


Figure 14: Illustration of bias vs variance

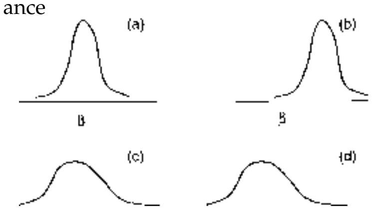


Figure 15: Another way to look at bias vs variance. a. is unbiased and efficient; b. is biased and efficient; c. is unbiased and inefficient; d. is biased and inefficient

8 Bootstrapping

Suppose you forgot how to calculate the standard error of the mean (the correct formula is $SE_{\bar{x}} = s/\sqrt{n}$), or another statistic. Or, more plausibly, you do not know the probability distribution of the data or the estimator. The bootstrap technique relies on the sample distribution instead to obtain an estimate. It draws a sample (with replacement) over and over from your sample and calculates a statistic from it. Using the distribution of these collected statistics (here the mean), we can calculate a confidence interval, for example.

For example:

```
setwd('~/Documents/Academia/Teaching/TCD/2015-HT/POTBD_Quantitative_Methods_II/Lectures/lecture1/')

#--- generate x ~ N[0,1]:
  x <- rnorm(10000, mean = 0, sd = 1)
#--- Calculate confidence interval for mean of x:
  # first, calculate estimated standard error of the mean
  esem <- sd(x)/sqrt(length(x))
  # calculate confidence interval
  ci.h <- mean(x) + 1.96*esem
  ci.l <- mean(x) - 1.96*esem

#--- Alternatively, using bootstrap:
  # sample over and over from x, and calculate the mean each time
  mean.x <- NULL
  for(i in 1:10000){
    x1 <- x[sample(1:length(x), size = length(x), replace = T)]
    mean.x <- c(mean.x,mean(x1))
  }

  mean.x <- mean.x[order(mean.x)]
  ci.boot.h <- mean.x[length(mean.x)*97.5/100]
  ci.boot.l <- mean.x[length(mean.x)*2.5/100]

#--- plot the bootstrap results
pdf('Figs/bootstrap.pdf')

  hist(mean.x, breaks=50)
  library(fields) # library to draw lines
  xline(ci.h, col=2, lty=2, lwd=2)
  xline(ci.l, col=2, lty=2, lwd=2)
  xline(ci.boot.h, col=4, lty=2, lwd=2)
  xline(ci.boot.l, col=4, lty=2, lwd=2)
  legend('topright', legend = c('estimated ci', 'bootstrap ci'), lty=c(2,2), col=c(2,4), cex=1)
dev.off()
```