Introduction to Statistics
the basics and beyond

Contents

Lets take the con out of econometrics

"Methodology, like sex, is better demonstrated than discussed, though often better anticipated than experienced."

(Leamer, 1983, page 40)

§0.1: Admin

These notes are building blocks. They are designed to help teach you how to apply statistical theory, build models, do proper data analysis, accurately estimate model parameters, make reasonable forecasts and run hypothesis tests **for the data and applications that interest you**.

The notes include examples. But to follow the examples, you *must* master some basic computer skills and statistical theory.

These are the best sources of basic statistical knowledge that I know.

- Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Penguin UK

- Freedman, D., R. Pisani, and R. Purves (2007). *Statistics* (4 ed.). Norton

- Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press

- James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. Springer

- Siegel, S. and N. J. Castellan (1988). *Nonparametric statistics for the behavioral sciences* (2 ed.). McGraw-Hill

The best book is Spiegelhalter (2019). The classic introductory text is Freedman et al. (2007).

0.1.3 Top 5: Advanced Text-Books

To go further, dip into the following list.

- Freedman, D. A. (2009). *Statistical models: theory and practice*. Cambridge University Press

- Efron, B. and T. Hastie (2016). *Computer age statistical inference*. Cambridge University Press

- Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (2 ed.). Springer

- Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and B. Rubin (2013). *Bayesian Data Analysis* (2 ed.). Chapman & Hall/CRC

- Casella, G. and R. L. Berger (2002). *Statistical inference* (2 ed.). Duxbury

The links between statistics and "machine learning" are found within Hastie et al. (2009) and Efron and Hastie (2016).

- **R:** `https://www.r-project.org/`

- **RStudio:** `https://www.rstudio.com/products/rstudio/download/`

- **HELP:** `https://stats.stackexchange.com` or `https://stackoverflow.com/`

- **SPSS:** `https://statistics.laerd.com/` for SPSS.

- **GraphPad:** `https://www.bioinformatics.babraham.ac.uk/training/GraphPadPrism/`

Investing time in R will save your time in the long-run. And it looks good on your CV, too. Also, R is free. SPSS and GraphPad Prism can be convenient in terms of "point-and-click" analysis; but they are both expensive and clumsy when it comes to repeating your analysis or using different data. GraphPad Prism makes excellent charts for small experiments, like mouse survival studies. SPSS links well with standard survey systems, such as Qualtrics.

Use books rather than on-line help when you can. Avoid Wikipedia.

A rough guide to the notation I use is as follows. There will be some small differences between these notes, articles and other books.

- Model variables are denoted by Latin characters a different font e.g. $y, x, Y, X \ldots$

- Unobserved "error" terms of models are usually denoted in Greek e.g. ϵ_i and ε_i; a model estimate of these is usually e_i.

- "True" model parameters are usually in Greek letters e.g. θ, α, β.

- Estimated model parameters models usually have a "hat" e.g. $\hat{\alpha}, \hat{\beta} \ldots$

- Matrices are usually large cap and bold e.g. $\boldsymbol{X}, \boldsymbol{Y} \ldots$ And vectors are usually bold and lower cap when not Greek e.g. $\boldsymbol{x}, \boldsymbol{y}, \boldsymbol{e}, \boldsymbol{\theta}, \boldsymbol{\beta}, \boldsymbol{\epsilon} \ldots$

- Expectation is denoted $\mathbb{E}[X]$ and other functions are written in normal font e.g. $\mathrm{var}(x), \mathrm{cov}(X, Y) \ldots$

French people adjust their language when "the rules" make something a bit ugly. Following this idea, I will sometimes break the conventions listed above.

Term	Meaning
IID	Independently and Identically Distributed
DGP	Data Generating Process
PDF	Probability Density Function
CDF	Cumulative Density Function
CLT	Central Limit Theorem
.....	.....

Chapter 1: Introduction

§1.1: Guiding Concepts

Section Goals ...

- Identify Fisher's key statistical tasks.
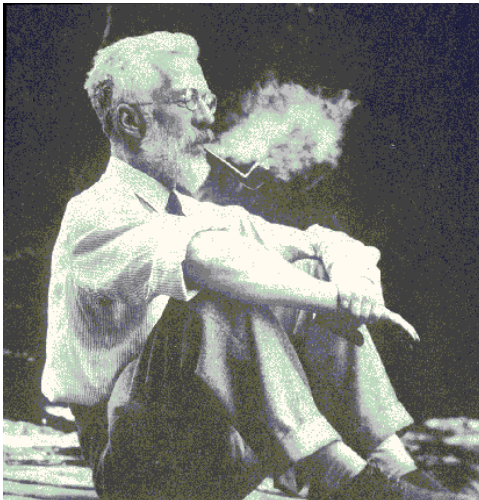
1.1.1 R A Fisher

R A Fisher was one the best statisticians ever. Among his accomplishments, he developed maximum likelihood estimation and founded modern hypothesis testing. He is also revered in genetics, in which he published over 100 scientific articles.

He thought that statistical analysis could not prove that smoking causes cancer.

To organise these notes and statistical thinking in general, we will use the three key statistical tasks that Fisher defined (Fisher, 1959, pages 6–8):

1. Model Specification.

2. Parameter Estimation.

3. Model Assessment (comparison and goodness-of-fit).

Statistical models can help answer questions about the real world. In broad terms, statistics is about:

- Specifying models designed to help us address interesting conceptual questions about return data.

- Using **finite samples** of data to estimate the parameters within these models.

- Assessing how well our models work — relative to the data we have and the conceptual questions that prompted us to build the models in the first place.

You should learn to define a "model" for every case in which you "do statistics" — so there is nothing mystical about what you do. But a persistent problem will be that we only ever have a limited sample of data with which we seek to address essentially **general conceptual questions** ...

§1.2: Sampling Variance

Section Goals ...

- Highlight sampling variance — perhaps the most important concept in statistics.

We always have to account for the error in the models we build that is due to using only a limited sample of data. We will call this "**sampling variance**".

But it doesn't matter what we call it — we all know what the problem is: any empirical estimates that we make could have been different if the sample had been different. The challenge is judging *how different* your estimates could have been.

Roughly speaking, accounting for sampling variance means coming up with a reasonable guess of the *variance* of the things we estimate.

For example, assume that you have estimated the mean return of a financial asset (or the change in baseline measure for a drug trial, or the survival time in a fruit-fly experiment, or whatever ...). The sampling variance of this estimate is the expected value of the square of the estimate less its true mean (i.e. $\mathbb{E}[(\hat{\mu} - \mu)^2]$), where the expectation is relative to the probability distribution of the data. The main problem is that we do not know the probability distribution of the data from which we took our **one-and-only** sample.

So sampling variances of estimates must be estimated themselves because we do not know the true distribution of the data. We will use three general methods.

- Direct estimates of variance based on **assumptions** about the data distribution.

- Approximate estimates of variance, using **plug-in** values and assumptions (also see the "delta method" in section 4.3.5).

- Non-parametric **bootstrap** estimates of variance or related quantiles.

The best method to tackle sampling variance will vary depending on the data and models involved. The maths behind the different techniques is not hard. The most difficult thing to grasp is the *concept* of sampling variance. We will return to the issue of sampling variance for each thing we empirically estimate.

Chapter 2: Models

This chapter specifies models that will enable us to kick-off any statistical analysis. By model specification, we mean to formally define and write a hypothetical structure.

"Everything is a model"?

§2.1: Model Specification

"Where do probability models come from? To judge by the resounding silence over this question on the part of most statisticians, it seems highly embarrassing. In general, the theoretician is happy to accept that his abstract probability triple 'Ω, A, P' was found under a gooseberry bush, while the applied statistician's model 'just growed'."

Dawid (1982)

Good models come from a variety of sources, including the following.

- Scientific theory.

- Creatively applied mathematics and statistics.

- Actual science and its practitioners.

- Suggested by the data itself (beware — see the sections on over-fitting and data-mining, in particular 2.2 on page 38).

We will not explicitly examine the source of models; but an aim is to enable everyone to confidently express their own ideas in terms of proper empirical models.

A good starting point is to recognise that our models will always be idealisations. It is a tautology to say that "all models are false, but some are useful" (attributed to George Box); but it is worthwhile considering the value of a "Galilean" approach to simplifications (allegedly, Galileo made many simplifications in order to make his calculations tractable, for example by assuming certain surfaces were frictionless).

We will use econometrics as an example. But what the econometricians say on the next few pages goes for any statistical pursuit.

Start somewhere

"Formulating a 'Good' Starting Point. We conceptualize the data-generating process (DGP) as the joint density of all the variables in the economy. It is impossible to accurately theorize about or precisely model such a high dimensional entity ... All empirical (and theoretical) researchers reduce the task to a manageable size by implicitly formulating a 'local DGP' (LDGP), which is the DGP in the space of the $m + 1$ variables $\mathbf{X}$ being modelled ... The choice of $\mathbf{X}$ is fundamental ..."

Hendry (2011, page 126)

We will use the term "data-generating-process" or "DGP" — dropping Hendry's 'local' — to loosely refer to any "true" form (i.e. one with yet-to-be-estimated parameters) of a model that we write down.

Do something, rather than nothing

"[There is a] a long tradition in econometrics of 'doing something without having to do everything.' This entails the study of partially specified models — that is, models in which only a subset of economic relations are formally delineated ... By allowing for partial specification, these methods gain a form of robustness. They are immune to mistakes in how one might fill out the complete specification of the underlying economic model"

Hansen (2013)

A general example is

$$\boldsymbol{y} = f(\boldsymbol{X}, \boldsymbol{\beta}) + \boldsymbol{\epsilon}, \tag{2.1}$$

where f can be any function of the data and the error term $\boldsymbol{\epsilon}$ can be given structural assumptions.

In reality we will make simplifying assumptions for f, such as assuming a linear relationship between the y-variable and x-variables. And we will specify convenient distributional assumptions for $\boldsymbol{\epsilon}$.

For a minimal DGP, you do not need to specify a particular probability distribution. Some interesting properties of the DGP will often be enough. The following is an example.

Let h stand for a DGP. Assume that each x_i, for $i = 1$ to N, is an independent random draw from h. And assume that h has a finite mean μ and finite variance σ^2.

In this minimal case, the central limit theorem (theorem 2 on page 132) applies. So, for example, the mean of samples from h will tend to be normal as the sample size N gets bigger. This DGP is therefore useful enough to do standard statistical testing.

Models have unknown *parameters*: these are what we estimate. Several examples follow.

- Assuming a particular probability distribution ("*pdf*") for data is a common basis for a "model". For example, if we assume that x is drawn from a Normally distributed population we need to estimate μ and σ from the following "model":

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}}\exp\left(\frac{-(x-\mu)^2}{2\sigma^2}\right) \tag{2.2}$$

- A linear DGP, which simplifies equation (2.1), can be expressed as

$$\boldsymbol{y} = \boldsymbol{X}\boldsymbol{\beta} + \boldsymbol{\epsilon} \tag{2.3}$$

  from which we seek to estimate $\boldsymbol{\beta}$.

- Further, we may assume a probability distribution applies to $\boldsymbol{\epsilon}$ in model 2.3; the parameters of this distribution will also typically require estimation (or assumed values).

The *data* enter models and estimation processes via the proverbial y-variables and x-variables from the models that we specify. For any particular model, how the data influences the parameter estimates depends on the estimation *method* used.

The sampling variance of the estimated parameters is a key ingredient in the assessment of models.

For example, let $\hat{\boldsymbol{\beta}}$ be the vector of parameter estimates from the model the equation 2.3. We require estimates for

$$\text{cov}(\hat{\boldsymbol{\beta}}) = \mathbb{E}[(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta})'], \tag{2.4}$$

from which the sampling variance of each element of $\hat{\boldsymbol{\beta}}$ is the corresponding diagonal element of $\text{cov}(\hat{\boldsymbol{\beta}})$.

The matrix notation is not important here. **What is important is the concept that the sampling variance of the parameters will in general depend on the sampling variance of data.**

Assume that the data set includes x_i for $i = 1$ to n. Let each x_i be IID with true mean μ and sd σ.

- The sample mean can be estimated as $\hat{\mu} = \sum x_i / n$.

- The sample variance can be estimated as $\hat{\sigma}^2 = \sum (x_i - \hat{\mu})^2 / (n-1)$.

A plug-in candidate for the *sampling variance* of the estimated mean is

$$\text{var}(\hat{\mu}) = \hat{\sigma}^2 / n$$

$$= \frac{1}{n} \sum_{i=1}^{n} (x_i - \frac{1}{n} \sum_{j=1}^{n} x_j)^2 / (n-1)$$

This is a simple function of the data.

Revealing simplifications

"In economics and as in other disciplines, models are intended to be revealing simplifications, and thus deliberately are not exact characterizations of reality; it is therefore specious to criticize economic models merely for being wrong. The important criticisms are whether our models are wrong in having missed something essential to the questions under consideration. Part of a meaningful quantitative analysis is to look at models and try to figure out their deficiencies and the ways in which they can be improved. A more subtle challenge for statistical methods is to explore systematically potential modelling errors in order to assess the quality of the model predictions."

Hansen (2013)

§2.2: Over-fitting

Section Goals ...

- Define over-fitting and multiplicity.

- Identify techniques that attempt to assuage the problems of over-fitting and multiplicity.

How did we find our best models? And why do they appear to fit in-sample data well? These two questions raise important issues for model assessment.

Some models come from years of painstaking academic research and publishing to-and-fro ("search and re-search"?). Others are derived by modern computing power which makes large scale model searches and estimation of complex models very easy. Further, modern research furnishes us with vast databases within which we can fish for apparently attractive models (Fischer Black called this type of data-mining "hindsight").

Let the cost to all the searching, path complexity, fishing, data-mining and inherent model flexibility be called "over-fitting". The consequence of over-fitting is that your model may explain very well the in-sample data, but it will not predict new data well. Said differently, over-fitting is a measure of how much of the *true* randomness from the in-sample data has been *erroneously* accounted for by an estimated model.

We can handle over-fitting if we can estimate it. This proves to be a challenging task. In this section we will investigate some limited methods to account for over-fitting.

Fischer Black was an entrepreneur, banker and academic. He co-authored the Black-Scholes formula. He used to eat cereal with orange juice rather than milk, but still died young.

Hindsight

"When a researcher tries many ways to do a study, including various combinations of explanatory factors, various periods, and various models, we often say he is 'data mining.' If he reports only the more successful runs, we have a hard time interpreting any statistical analysis he does. We worry that he selected, from the many models tried, only the ones that seem to support his conclusions. With enough data mining, all the results that seem significant could be just accidental. ... Less formally, we call it 'hindsight'."

Black (1993)

Over-fitting is a insidious problem; there are often complex motives for searching across many model structures, methods and data sets.

The following general issues should be kept in mind.

1. The more flexible an *overall* model selection or search procedure is the more inaccurate truly-out-of-sample predictions will be. This is Fischer Black's "**hindsight**". Unfortunately, it is difficult to accurately measure the affect on over-fitting from trying out many models.

2. The *types* of models which make good in-sample predictions typically have lower apparent errors at the cost of higher out-of-sample errors.

3. Over-fitting is a dance between the data and the structure of a model — whilst some models are *in general* more flexible than others, over-fitting is relative to the data used in any given case.

Chapter 3: Inference

§3.1: Hypothesis Testing

Section Goals ...

- Provide background reading for class discussion.

- Clarify the philosophical reasoning behind statistical tests.

- Define a "p-value" and "p_0-value".

This section first discusses the general nature and background of statistical tests. Then so-called "p-values" and "p_0-values" are defined. The concept of confidence intervals is then analysed using p_0-values.
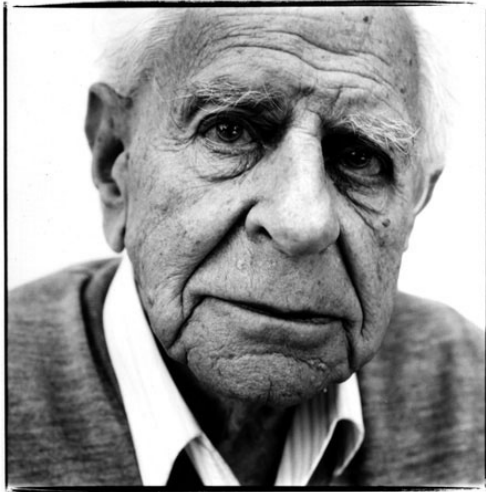
Karl Popper was a philosopher of science in the 20th century. Popper's main theory about scientific methodology is that we advance knowledge by "conjecturing" theories, which we subsequently try to "refute". This view is not widely accepted today — in particular because there is a paucity of "crucial experiments" which definitively refute theories.

Indeed, we cannot do much real "refuting" in real science or statistics. But we do form substantive null hypotheses and try to *"test"* them. In this context, we have good Popper-like principles to follow:

- Be explicit about our premisses for models and hypotheses.

- Do as many tests on your models and hypotheses as you can (be reasonable).

- Present *all* your key statistical results. And do this in a format that is useful for others.

- Be creative and critical with setting up models and forming hypotheses.

In each test we do we will specify a "null hypothesis" (or simply "null", if you like).

- The null should be something useful and interesting, like a sort of benchmark. For example:
    - "all the parameters in the model are zero"
    - "the slope co-efficient in the linear regression is zero"
    - "the true means of the distributions for x and y are the same"

- The null must be precise — something like "model A fits the data better than model B" is too vague.

- For each null we should try to think of rival alternative. Usually this is straight-forward. But not always.

The idea for each test is to calculate a measure, usually a p-value, with which the null hypothesis can be judged. We will define p-values and the process of judgement on the next few pages.

R. A. Fisher proffered a statistical method of Popper's *refutation*: null hypothesis test "rejection".

- Fisher says that we "reject the null" when the probability of the evidence, given the null is true, falls in an "extreme" region of its probability distribution.

- In other words: Fisher tells you to reject the null if you were extremely unlikely to observe the data that you did, *if* the null was in fact true.

- Fisher's intention was *not* some behavioural response – his idea of rejection was "logical".

Neyman and Pearson added to Fisher's work. They insist that a null hypothesis be tested against an *explicit alternative*.

- The two approaches are basically the same, if you trust Fisher's selection of "extreme regions".

- But a Neyman-Pearson "rejection" has a more natural behavioural interpretation: it would influence what you do in *repeated* experiments. Fisher thought this was un-scientific – partly because we often want to do testing in one-off situations with a unique set of data.

3.1.5 Errors of Decision: I and II

We will put to one side the philosophical concerns about the meaning of "rejecting" or "accepting" a null hypothesis.

If you decide to accept or reject a hypothesis then you leave open the possibility of error, relative to the truth. These errors are conventionally called "Type I" error and "Type II" error.

	H_0 **is True**	H_0 **is False**
Accept H_0	✓	Type II error
Reject H_0	Type I error	✓

Thomas Bayes gave us the formula:

$$\mathbb{P}(H|E) = \frac{\mathbb{P}(H)\mathbb{P}(E|H)}{\mathbb{P}(E)},$$

where H is the hypothesis, E is the evidence (or "data", in some sense), $\mathbb{P}(H)$ is the "prior" probability of the hypothesis being true and $\mathbb{P}(E|H)$ is the probability of the evidence *given* the hypothesis is true. The portrait shown to the left is probably of Thomas Bayes ...

1. The Bayesian view is good for "decision science". But its not so good if you do not have explicit prior probabilities. And we usually don't.

2. So we typically "model" the likelihood: $\mathbb{P}(E|H)$. This is R A Fisher's focus!

3. We even have tests based on likelihood ratios e.g. $\mathbb{P}(E|H_0)/\mathbb{P}(E|H_a)$

We will NOT calculate the probability of a hypothesis being true or not.

The following is a recipe that results in what we will call a "p$_0$-value" (if you can calculate a p$_0$-value then a corresponding p-value follows directly).

1. Cook-up a model.

2. Measure or collect some related data.

3. Estimate the parameters of the model with the data.

4. Cook-up an interesting hypothesis.

5. Turn the data-model-hypothesis combination into "evidence" about the hypothesis.
   – This requires calculating a "test statistic" with a *known probability distribution* when the null hypothesis is true.

6. Let the p$_0$-value be the test statistic's value in the cumulative density function of the test statistic "under the null" — call this value p_0.

A p$_0$-value ("p_0") is calculated by looking up the value of the test statistic in its applicable *cdf*.

- Both p$_0$-values and p-values look up the *cdf* of the test statistic. The difference between a p$_0$-value and corresponding conventional p-value depends on the specific test. For example, in a typical t-test the p-value looks up the *cdf* value of the absolute value of the test statistic.

- The *cdf* look-up process renders all p$_0$-values to be distributed as $U(0,1)$ random variables when the null hypothesis is true (or "under the null"). The same is true for conventional p-values.

- The p$_0$-value indicates interesting features about the data and hypothesis (e.g. whether a mean appears high or low relative to the conjectured null value). Some of this information can be lost in conventional p-values.

- When combining many different hypothesis test results p$_0$-values are more convenient than p-values; for example, with false discovery analysis (see section 4.2 on page 118).

Conventional p-values, call them p_{con}, are defined as a transformation of the value of a test statistic in the applicable *cdf*. The transformation depends on the specific details of the hypothesis test. A general recipe for finding *conventional* p-values is as follows.

- If the hypothesis test is "two-sided" (e.g. when the null is $H_0 : \mu = 0$ and the alternative is $H_a : \mu \neq 0$) then $p_{con} = 2(1 - p_{abs})$, where p_{abs} is the value of the *absolute value* of a test statistic in its applicable *cdf* (when the null hypothesis is true).

- If the alternative hypothesis is "one-sided" (e.g. when the null is $H_0 : \mu > 0$ and the alternative is $H_a : \mu \leq 0$) then $p_{con} = (1 - p_{abs})$.

- For the special case in which the alternative hypothesis implies that evidence against the null hypothesis comes exclusively from test statistics that are extremely *positive* then $p_{con} = 1 - p_0$. Note that in these cases, we typically have $p^{abs} = p_0$ because the test statistic itself is a quadratic — see section 3.1.10 on page 54.

Test statistics in "quadratic" form typically imply evidence against the null when p_0 is very high. In this case, p-values are conventionally reported as $p_{con} = 1 - p_0$. As in other cases, both p_0 and p_{con} are distributed $U(0, 1)$ under the null. Quadratic form tests usually involve test statistics that follow the χ^2 or F distributions, under the null. A defining feature of quadratic form tests is that the p-value does not identify "the direction" from which the data are inconsistent with the null hypothesis, though in some examples this can easily be recovered.

For example, assume a t-test statistic $t_x \overset{d}{\to} N(0, 1)$ under the null hypothesis that "the true mean of x is 0" (versus the alternative hypothesis that "the mean of $x \neq 0$"). Then we also have $t_x^2 \overset{d}{\to} \chi^2(1)$, which is in quadratic form. So, assuming the sample size is adequate, we have the option of comparing the square of the test statistic with a chi-squared distribution. The t-test p_0-value will be more *directly* informative with respect to violations of the null — we would get an indication of whether the mean was more likely to be above or below the zero (this is not true of p_{con}).

However, the quadratic forms common in testing typically invoke compound or *joint* hypotheses about multiple parameters. In these cases, "the direction of null hypothesis violation" is not of primary importance or may not even make sense.

In traditional statistical literature, p_{con} is the standard definition of a p-value. And with p_{con} in hand, the traditional process is to "reject the null" if p_{con} is below some extreme cut-off "α". A typical cut-off is $\alpha = 0.05$. In this case, if a conventional p-value is less than 0.05 then there is less than a five percent chance of observing such an extreme p-value under the condition that the null hypothesis is true.

We can use a p_0-value in similar fashion. This means looking for one of the following conditions, depending on the hypothesis test at hand:

- $p_0 < .05$;

- $p_0 > .95$; or

- $p_0 < .025$ and $p_0 > .0975$.

However, the idea behind p_0-values is to report them explicitly, rather than relying on rejection rules and arbitrary cut-off regions.

A simple example of how to generate p$_0$-values is as follows.

- Assume you have gathered a sample of n asset returns which has a sample (or observed) mean of $\bar{x}$ and unknown true mean μ.

- Assume each x_i is independent and identically distributed (IID) to all the others.

- You want to test a series of J null hypotheses that the true mean of the asset returns is equal to μ_j for $j = 1$ to J.

- The test statistic can be the difference between the observed mean and the assumed-to-be-true mean, divided by the square root of the sampling variance of the observed mean.

- Let p_0^j denote the p$_0$-value for null hypothesis j.

Continued on next page ...

So we have p_0^j, corresponding to different "conjectured true means', μ_j:

$$p_0^j \;=\; \mathbb{P}\left(Y \le \frac{\bar{x} - \mu_j}{\hat{\sigma}/\sqrt{n}} \right) \tag{3.1}$$

where it is reasonable to assume $Y \sim T(n-1)$, $j = 1$ to J and $\hat{\sigma}$ is the sample standard deviation of x. And we know that we can also assume that $p_0^j \sim U(0,1)$ if $\mu = \mu_j$. The conventional p-value corresponding to equation (3.1) is

$$p_{con}^j \;=\; 2\left(1 - \mathbb{P}\left(Y \le \left| \frac{\bar{x} - \mu_j}{\hat{\sigma}/\sqrt{n}} \right| \right) \right). \tag{3.2}$$

This will also have a $U(0,1)$ distribution under the null (and associated assumptions).

Most of the parametric hypothesis testing (i.e. using analytic probability distributions) in statistics comprises the following three components.

1. An estimated value which is in some sense is a sample *mean*.

2. The true value (as *assumed* by a null hypothesis).

3. The sampling variation (or its square root) of the estimated value.

Typically the values in 1, 2 and 3 can be combined in such a way that a known probability distribution applies. The link to a distribution often invokes the Central Limit Theorem applies (this will often apply, in some form, since we are dealing with averages).

For example, the test statistic in equation (3.3) on page 59 is akin to a simple function of $\frac{(1)-(2)}{(3)}$, which is T-distributed. Most other sophisticated test statistics boil-down to combinations of these same three components.

Assume the same set-up and assumptions that lead up to equation (3.1). Let the "test statistic" for a sample mean $\bar{x}^*$ (here considered to be a random variable) be

$$t^* = \frac{\bar{x}^* - \mu}{\hat{\sigma}/\sqrt{n}}, \tag{3.3}$$

where μ is the true but unknown mean of x. And let the inverse *cdf* value from a T-distribution with $n - 1$ degrees of freedom at the quantile α be $T_\alpha^{-(n-1)}$. Now we have

$$\mathbb{P}\left(T_{\alpha/2}^{-(n-1)} \leq t^* \leq T_{1-\alpha/2}^{-(n-1)}\right) = 1 - \alpha \tag{3.4}$$

Equation (3.4) is related to a so-called "confidence interval" for a *particular* estimate $\bar{x}$ which is ordinarily defined as the interval

$$\left[\bar{x} - T_{1-\alpha/2}^{-(n-1)}(\hat{\sigma}/\sqrt{n}) \, , \, \bar{x} - T_{\alpha/2}^{-(n-1)}(\hat{\sigma}/\sqrt{n})\right], \tag{3.5}$$

where $\bar{x}$ is a single-case "random draw" from all the possible $\bar{x}^*$ values and α is typically 0.05.

The interval in (3.5) is derived in two steps. First, re-arranging the inequality in the left-hand-side of equation (3.4) so that μ appears in the middle. Secondly, plugging-in an observed $\bar{x}$ in place of the random variable $\bar{x}^*$. **The interval in (3.5) is NOT a simple probabilistic statement about the true mean of x.**

The probability in equation (3.4) is a statement in which $\bar{x}^*$ is a *random variable* — so the probability relationship does not transfer to the confidence interval in (3.5) because the interval is a statement about the plugged-in mean of an actual sample. For the avoidance of doubt: the confidence interval does **not** say that the true mean of the underlying distribution for x lies within the interval with probability 95%. That would in fact be a Bayesian probability statement, which requires a prior probability distribution for the true mean!

There is a direct relationship between the interval in (3.5) and the p_0-values from formula (3.1). The confidence interval is defined by the values $[\mu_{low}, \mu_{up}]$ such that

$$p_0^{up} = \mathbb{P}\left(Y \le \frac{\bar{x} - \mu_{up}}{\hat{\sigma}/\sqrt{n}}\right) = \frac{\alpha}{2}, \tag{3.6}$$

$$p_0^{low} = \mathbb{P}\left(Y \le \frac{\bar{x} - \mu_{low}}{\hat{\sigma}/\sqrt{n}}\right) = 1 - \frac{\alpha}{2}. \tag{3.7}$$

In other words, the confidence interval corresponds to p_0-values where we *assume* that the true mean is μ_{low} and μ_{up}.

So calculating a series of p$_0$-values from a wide enough range of assumed-to-be-true values is more informative than calculating a confidence interval: you get the confidence interval bounds *and* everything in between.

In the context of the t-test in equation (3.1) this would entail a series of tests with varying levels of μ_j.

Loosely speaking, the "power" of a test set-up is its ability to appropriately produce extreme p-values when the null hypothesis is *false*. The key to power is test **sensitivity** to violations of the null.

The flip-side of power is the tendency to produce extreme p_0-values when the null hypothesis is *true*. The "test is correctly specified" if under the null the p_0-values are $U(0, 1)$ distribution. The key to correct specification is having the right distributional assumptions about the test statistic.

The "size" of a test is the probability that an extreme p_0-value occurs when the null hypothesis is true. If the test is correctly specified then the size of the test is a direct function of what you regard as extreme. For example, if you regard as extreme p_0-values outside the interval [.025,.975] then the size of the test will be 5%.
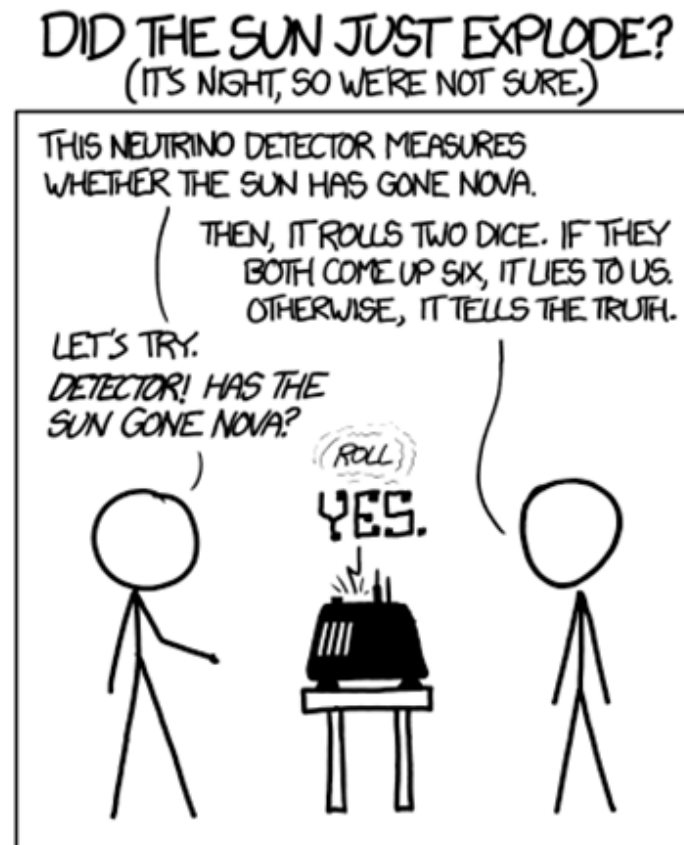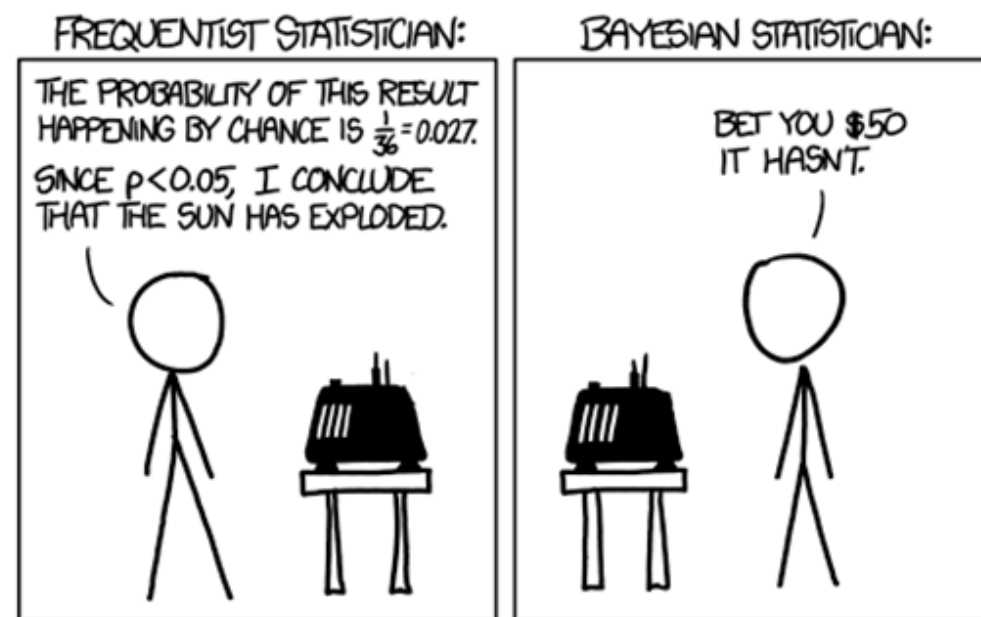
§3.2: Testing Examples

Section Goals ...

- Examine basic examples of statistical tests.

Assume you have a vector of sample data $\boldsymbol{x}$ of length n and you want to test the null hypothesis be "$H_0 : \mu = 0$" versus the alternative "$H_a \neq 0$". The conventional approach is conduct a classical "t-test".

The first calculation step is to estimate the mean and variance of x. Assume for simplicity that the data are independent. Now assume that the true data distribution is "well behaved enough" that samples of length n are sufficient to ensure sample averages are approximately Normal and standard deviations can be estimated appropriately. We can then form a classical t-statistic, call it t_x here, under the conditions implied by the null hypothesis.

In other words, we have (approximately, at the least)

$$t_x = \frac{\sqrt{n}\hat{\mu}}{\hat{\sigma}} \sim T(n-1), \tag{3.8}$$

if the null hypothesis is true.

The associated p-value is straight-forward to calculate: just look up the cdf for the t-distribution at the point t_x.

3.2.1 Individual Mean Tests

Exercise 1 (A simple t-test?).

A. *Derive the formula for a conventional t-test from first principles. State what the key assumptions are. You can assume that σ is known and start from the central limit theorem (making it a "z-test" rather than a t-test).*

B. *If the data are correlated how can the t-test take this into account?*

	n	mean	sd	t_x	p_0-value	$CI_{0.025}$	$CI_{0.975}$
S&PCOMP	222	0.033	0.505	0.981	0.836	-0.034	0.100
@AAPL	222	0.212	1.506	2.096	0.981	0.013	0.411
U:XOM	222	0.039	0.588	0.985	0.837	-0.039	0.117
U:KO	222	0.022	0.595	0.554	0.710	-0.057	0.101
NASCOMP	222	0.033	0.780	0.633	0.736	-0.070	0.136
FundA	222	0.042	0.539	1.156	0.876	-0.029	0.113
FundB	222	0.064	0.939	1.020	0.846	-0.060	0.189
FundC	222	0.035	0.350	1.473	0.929	-0.012	0.081
FundD	222	0.051	0.445	1.695	0.954	-0.008	0.109

Table 3.1: Results from t-tests on monthly asset returns. The null hypothesis in each case is that the mean monthly return is zero. Monthly asset returns have been annualised (multiplied by 12). CI_x is the confidence interval cut-off at quantile x.

	n	mean	sd	t_x	p_0-value	$CI_{0.025}$	$CI_{0.975}$
@AAPL - S&PCOMP	222	0.179	1.337	1.991	0.976	0.002	0.355
U:XOM - S&PCOMP	222	0.006	0.566	0.149	0.559	-0.069	0.081
U:KO - S&PCOMP	222	-0.011	0.611	-0.271	0.393	-0.092	0.070
NASCOMP - S&PCOMP	222	0.000	0.447	-0.004	0.498	-0.059	0.059
FundA - S&PCOMP	222	0.009	0.181	0.705	0.759	-0.015	0.033
FundB - S&PCOMP	222	0.031	1.129	0.409	0.659	-0.118	0.180
FundC - S&PCOMP	222	0.001	0.359	0.056	0.522	-0.046	0.049
FundD - S&PCOMP	222	0.017	0.631	0.409	0.659	-0.066	0.101

Table 3.2: Results from *paired* t-tests monthly asset returns less market index returns. The null hypothesis in each case is that the mean monthly return in excess of the index is zero. Monthly returns have been annualised (multiplied by 12). CI_x is the confidence interval cut-off at quantile x.

Exercise 2 (T-test interpretation).

    A. *Carefully interpret the results of the last two t-test tables.*

    B. *What inferential issues can you imagine if we ran 2000 paired t-tests (versus the index return) for each of the largest 2000 stocks in the US?*

    C. *Would it be worthwhile splitting the data in different time periods and performing separate tests? What are the issues with doing this?*

§3.3: Bootstrapping

Section Goals ...

- Introduce bootstrapping.

Classical tests, like t-tests and Wald-style tests, are parametric in the sense that an analytic distribution is *assumed* to hold under the null hypothesis. There can be significant problems with this approach, including the scenarios described below.

1. The data distribution is unusual in the sense that finite sample test statistics are not even close to their assumed distributions. This problem can occur due to high levels of skewness, the presence of outliers and fat-tailed data.

2. The estimate is such a complicated function of the data that the sampling variance is not tractable. For example, see the return measures like the Treynor ratio.

3. You are not even sure what test statistic or asymptotic distribution should be used. A related problem is when the first order Taylor series for an estimating function has significant second and third order terms (see section 4.3.5 on page 134).

A potential solution to these problems is called "**bootstrapping**". This is a non-parametric approach in the sense that no analytic distribution is assumed to apply. But you still need assumptions ...

Bootstrapping is based on the idea of *re-sampling* the data that you actually have. Each new sample is referred to as a "bootstrap sample". And for each bootstrap sample, you re-calculate the statistic in which you are interested. The resulting *distribution* of statistics (across bootstraps) more-or-less represents all the statistics that *"you could have got"*: in other words, the distribution allows you to estimate the sampling variance of the statistic that you actually did get.

This sounds like magic, and perhaps must be seen to be believed. A rigorous introduction is found in Efron, B. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Gentle introductions to bootstrapping include Shalizi (2010) and Hesterberg (2015).

3.3.2 Bootstrapping

Several assumptions must hold for bootstrapping to be worthwhile.

- The actual sample must be "representative" of the data population.
    - Importantly, this requires enough data from the extreme tails of distributions to be present.

- The re-sampling must be appropriately random — usually this means re-sampling the actual sample data points with equal probability and replacement; with each bootstrap sample being the same size as the original sample.

- The number of bootstrap samples needs to be "high enough" — especially to incorporate enough of the extreme tail observations.

Bootstrapping is a powerful and general method with which sampling variance can be estimated! And we do not have to be mathematical geniuses. We just need to re-sample our original data many times.

Many extensions to bootstrapping have been developed.

- **Block bootstrapping**: when data are suspected of having time series dependence then *blocks* of data should be re-sampled together. For example, if weekly asset returns have two lags of positive correlation then the block size for each re-sample should be at least three contiguous weeks.

- **Parametric bootstrapping**: sometimes it is reasonable to *assume* the data (or part of it) follows a particular analytic distribution. In these cases, you can arguably re-sample from the assumed distribution rather than the original data.

3.3.5 Bootstrap Confidence Intervals §Bootstrapping

It is straight-forward to bootstrap confidence intervals for virtually any estimate. A basic recipe is as follows.

1. Re-sample the data B times and recalculate the estimate $\hat{\gamma}$ each time.

2. Order the B values of $\hat{\gamma}$ from smallest to largest into a "bootstrap cumulative distribution" $F_b(\hat{\gamma})$.

3. Find the cut-off points in $F_b(\hat{\gamma})$ that you desire for a $(1 - \alpha)$ based confidence interval.

The "devil is in the detail" — for example, what should the cut-off points be?

Exercise 3 (Bootstrap cut-offs). *Why is setting cut-off values a problem?*

Another option is to construct an interval based on the quantile cut-offs of a bootstrap test-statistic distribution (see the "bootstrap t" example in Hesterberg, 2015). This may be more accurate for small non-normal datasets.

3.3.6 Bootstrap Testing and P-values

Bootstrapping can be used to test hypotheses as well. The best recipe will depend on the specific form of the null hypothesis.

- A popular approach requires the bootstrap samples to be consistent with the null hypothesis. Exactly the same idea is used in parametric testing. The trick is choosing how the null hypothesis is "enforced" upon bootstrap samples — there can be subtly different options. In this case, a test's p_0-value could be represented by the quantile at which the *actual* sample value lies amongst the corresponding bootstrap values, which are consistent with the null being true.

- Another option is to calculate a bootstrap confidence interval for the estimate and "reject" any hypothesis that posits a true value outside the interval.
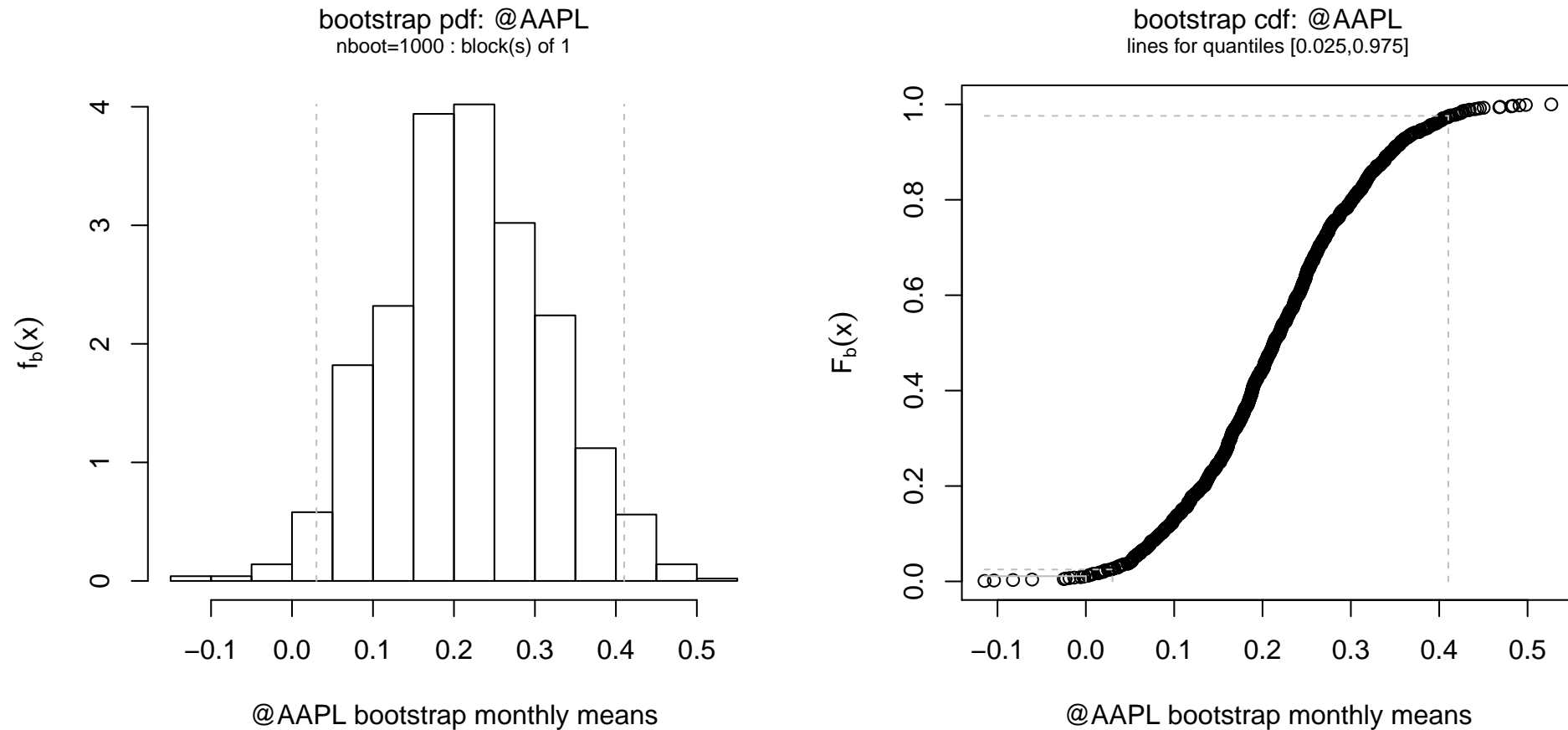
The following pages present empirical distributions for bootstrap samples of returns for Apple and the S&P 500 index. In each case, the same time periods were re-sampled. Note that a "Sharpe Ratio" is the excess return of an asset return with respect to a risk free asset (the cash rate, if you like), divided by the standard deviation of the asset's returns.
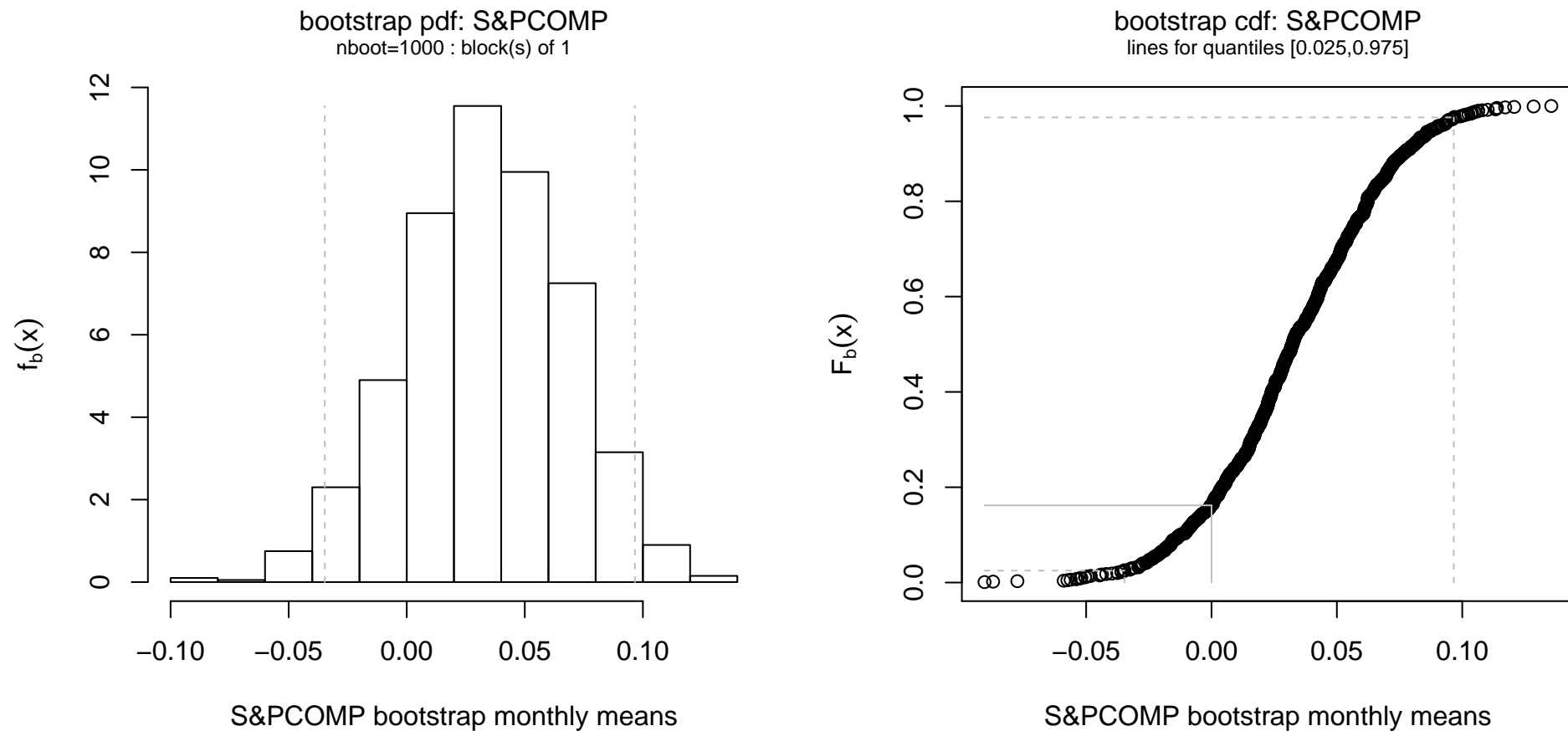
bootstrap pdf: @AAPL
nboot=1000 : block(s) of 1

bootstrap cdf: @AAPL
lines for quantiles [0.025,0.975]

@AAPL bootstrap monthly means

@AAPL bootstrap monthly means

Figure 3.1: Bootstrap distribution of the mean for monthly Apple returns, since 2000.

Figure 3.2: Bootstrap distribution of the mean for monthly S&P 500 returns, since 2000.

bootstrap pdf: S&PCOMP−@AAPL
nboot=1000 : block(s) of 1

S&PCOMP−@AAPL bootstrap diff in sharpe ratios

bootstrap cdf: S&PCOMP−@AAPL
lines for quantiles [0.025,0.975]

S&PCOMP−@AAPL bootstrap diff in sharpe ratios
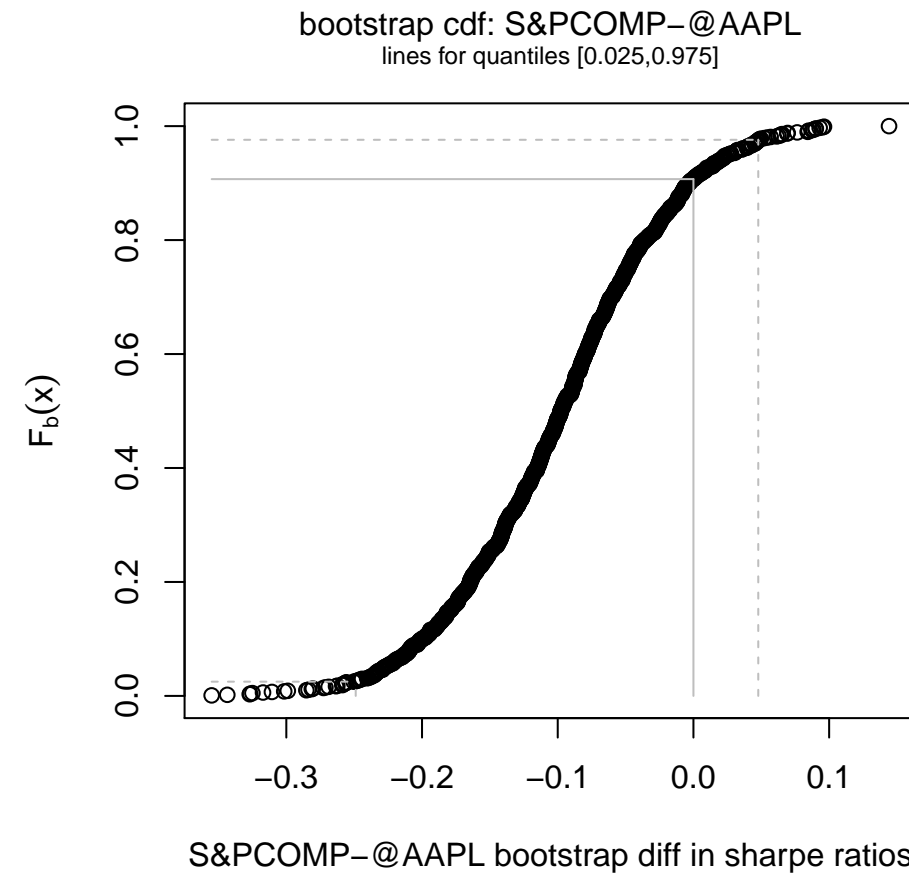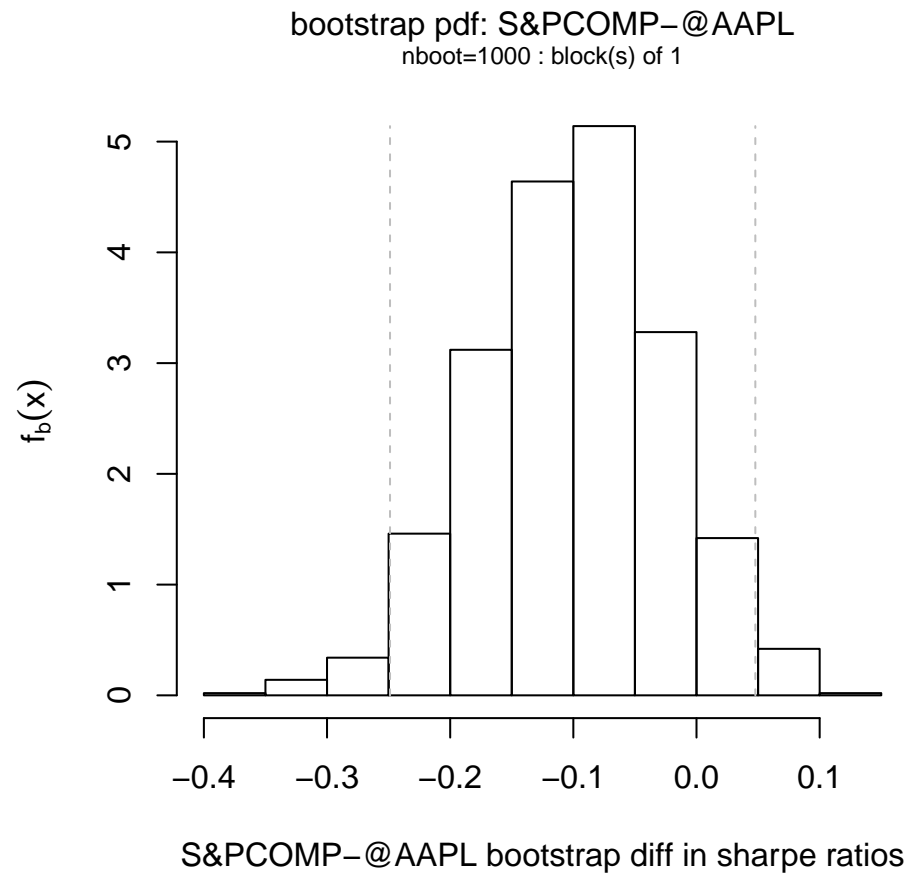
Figure 3.3: Empirical *pdf* and *cdf* for bootstrap Sharpe Ratio differences for S&PCOMP-@AAPL. Monthly data from 2000-01-31 to 2018-06-29.

Exercise 4 (Bootstrapping).

A. *How do the bootstrap results in figure 3.1 and figure 3.2 compare to the t-tests in table 3.1 on page 67?*

B. *How can you construct evidence to test whether Apple has a higher average return than other assets?*

§3.4: Questions to Answer

3.4.1 Reminder & Prelude

- We are not doing **mathematical statistics**.

- We have computers. We have theory. And we have lots of data. But we need to use these things wisely.

- We should aim for useful and wise inferences. *"Inference: 1.a The action or process of inferring; the drawing of a conclusion from known or assumed facts or statements; "* OED.

4.2.1 Models.

A. Give an example of a simple data-generating-process on which you could base a useful model.

B. Write down two different models of any sort and identify the parameters and the variables.

C. Give an example of how a model enables a meaningful statistical test.

4.2.2 Hypothesis Tests.

A. What is the point of a hypothesis test?

B. Describe the properties of null hypotheses.

C. Give an example of an ill-formed null hypothesis and say what is wrong with it.

D. When is a hypothesis test decisive (or directly useful for action, if you like)?

4.2.3 Test statistics.

A. What is the role of a "test statistic" in a hypothesis test?

B. What statistical elements do most test statistics have in common?

C. Let $t^* = \frac{\bar{x}^* - \mu_0}{\hat{\sigma}/\sqrt{n}}$. Assuming that the true mean μ is not equal to μ_0, what things will tend to make t^* have a more extreme p-value?

4.2.4 P-values.

A. Define in your own words what a p-value *means*.

B. What is a common misinterpretation of a p-value?

C. Describe a sufficient number of steps to produce a p-value (in a simple context).

D. What distribution does a p-value follow when the null hypothesis is true?

E. Is the concept of an "extreme" p-value subjective or objective?

4.2.5 Confidence Intervals.

A. Describe how confidence intervals are related to p-values (or p_0-values).

B. Is a confidence interval more or less informative than a series of p_0-values for null hypotheses that span a wide range of posited true parameter values? Why?

4.2.6 Test power.

A. What is the "size" and the "power" of a hypothesis test?

B. Define Type I error and Type II error.

C. How can power for a particular test be investigated?

§3.5: Hypothesis Test Tutorial

Section Goals ...
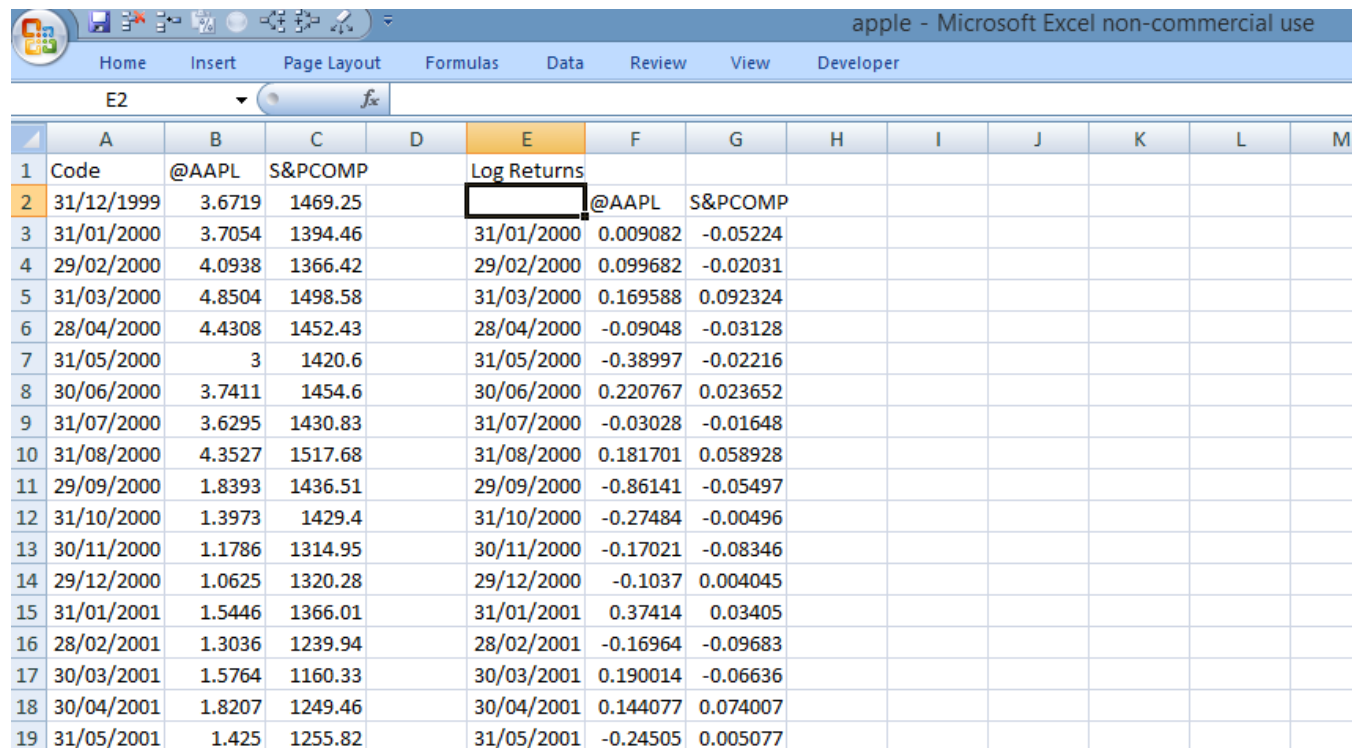
- Examine an indicative test set-up and associated inference errors.

- Explicate size and power.

- Run real tests on a computer.

Assume we have a sample vector $\boldsymbol{x}$ of length n. We will refer to each element of $\boldsymbol{x}$ as x_i, for $i = 1, 2, ..., n$.

To make it concrete, assume that each x_i is the monthly log return on Apple shares less the log return on the S&P 500 index. Call this series the "excess returns" for Apple.

Assume that we are interested in whether the average return for Apple in each period is different from matching return from the index.

To prevent this question being solely about a historical fact, we will assume that each x_i is in some sense generated by a random process and call this process "the model". Taking that leap, we can make assumptions about the model, estimate parameters for it and run statistical tests.

Assume a minimal DGP.

Let h stand for the DGP which generates Apple returns. Assume that each x_i for $i = 1$ to n is a random draw from h and each x_i is independent of the others. Assume that h has a finite mean μ and finite variance σ^2.

Assuming the DGP h holds, we can formalise a hypothesis based on our question about the average excess returns for Apple.

The question implies the formal null hypothesis that

$$H_0 : \mu = 0. \tag{3.9}$$

Given the assumptions about h and the null hypothesis in (3.9) we can utilise the proverbial t-test. This entails the following.

First, we calculate the applicable test statistic

$$t_x = \frac{\hat{\mu} - \mu_0}{\hat{\sigma}/\sqrt{n}},$$
(3.10)

in which $\hat{\mu}$ is the average of the sample $\boldsymbol{x}$, $\mu_0 = 0$ and $\hat{\sigma}$ is the sample standard deviation.

Under the null hypothesis (and model and assumptions), t_x can be assumed to be t-distributed with $n-1$ degrees of freedom. In other words, we assume that

$$t_x \sim T(n-1)$$
(3.11)

under the null.

So far we have

$$H_0 : \mu = 0$$

and

$$t_x = \frac{\hat{\mu} - 0}{\hat{\sigma}/\sqrt{n}} \sim T(n - 1),$$

under the null hypothesis and its related assumptions.

Discussion

1. What can we do with this? [hint: see section 3.1.7 on page 51]

2. What is missing?
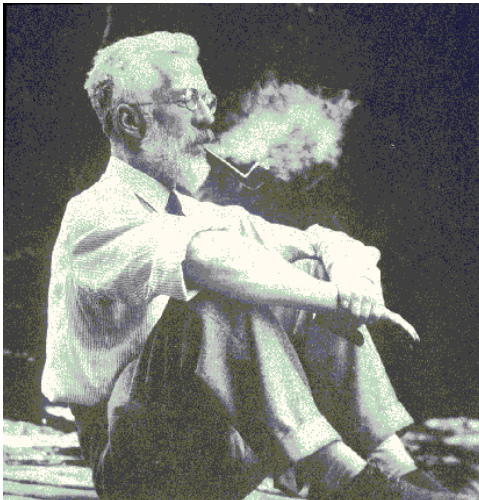
R A Fisher was one the best statisticians ever. Among his accomplishments, he developed maximum likelihood estimation and founded modern hypothesis testing. He is also revered in genetics, in which he published over 100 scientific articles.

He thought that statistical analysis could not prove that smoking causes cancer.

We have $H_0 : \mu = 0$ and $t_x = \frac{\hat{\mu} - 0}{\hat{\sigma}/\sqrt{n}} \sim T(n-1)$. Now we look up the value of the cumulative density function of the $T(n-1)$ distribution at the point t_x. Assume t_x is in the extremes (or "tails") of the *cdf*. Fisher would infer that:

"either the null hypothesis is true, in which case something unusual happened by chance, or the null hypothesis is false."

> **Question**
> 1. Is this inference satisfying?

"either the null hypothesis is true, in which case something unusual happened by chance, or the null hypothesis is false."
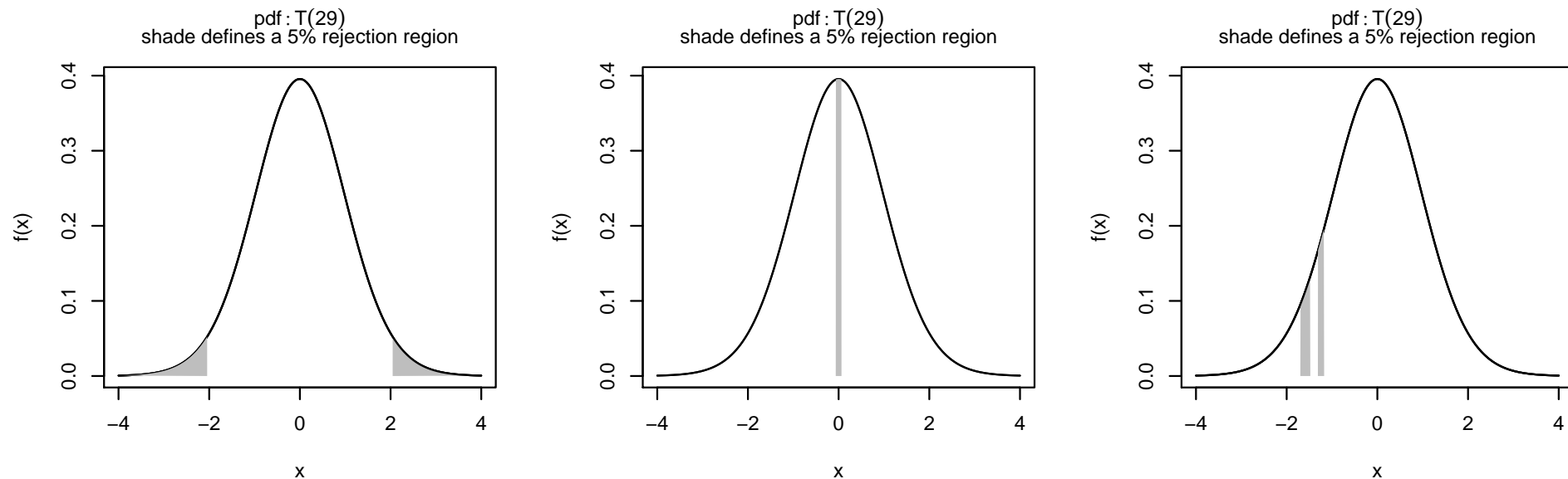
Figure 3.4: Different options for 5% rejection regions.

The key problem with Fisher's inference is that the rejection region is not pinned down — it could be anywhere and the inference still follows.

We can *technically* augment Fisher's bare inference by adding an explicit alternative hypothesis and a new "decision process" about rejecting the null hypothesis, *relative* to the alternative.

Let the alternative hypothesis in our case be conditional catch-all

$$H_a : \mu \neq 0, \tag{3.12}$$

which makes the overall set-up a "composite" hypothesis test.

Assume that the under the alternative hypothesis the same model applies and that the same assumptions are as before, except that one of the model's parameters is different.

Questions

1. Does defining extreme to be in the tail ends of the t-distribution make sense now?
2. Consider "simple" alternative hypotheses like $H_a : \mu = 2$. And then think about a class of simple alternatives that would define $H_a : \mu \neq 0$.
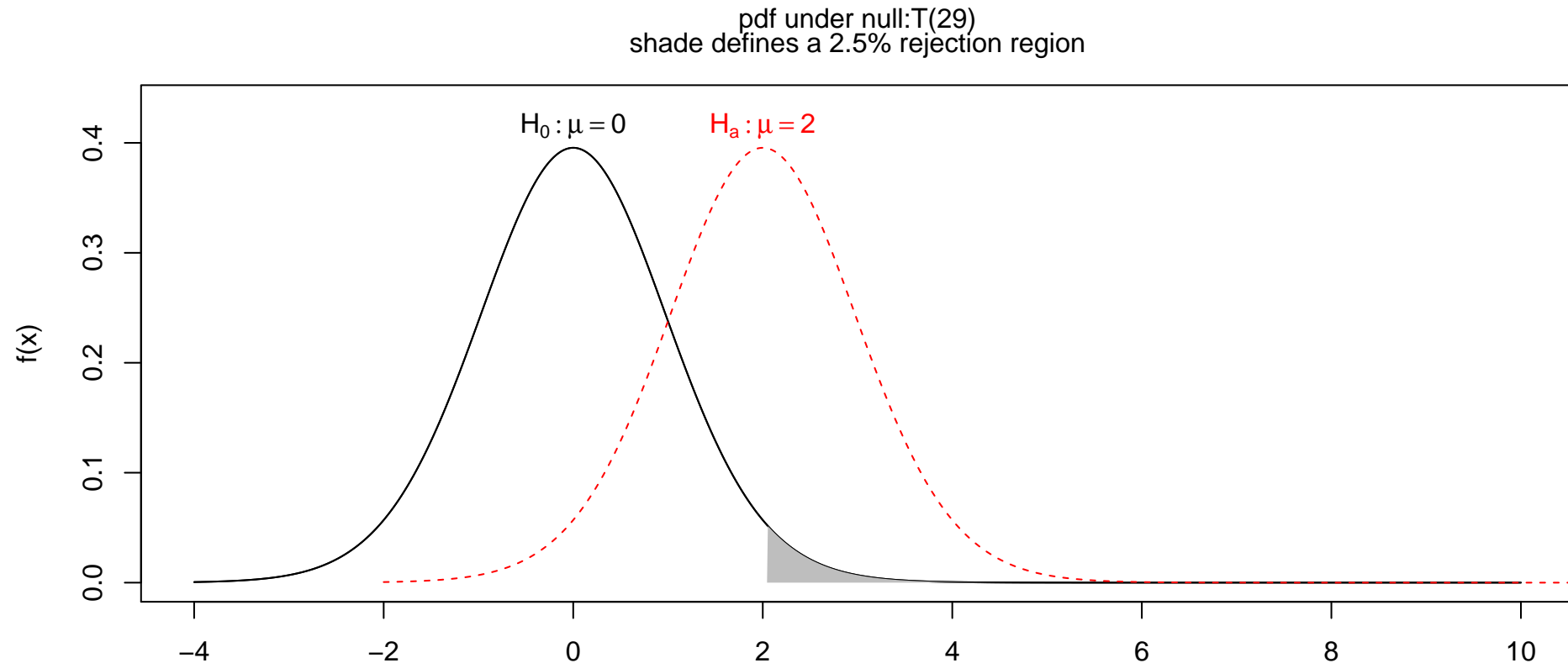
Figure 3.5: A simple alternative to the null, which is t-distributed.

With reference to figure 3.5, let $H_0 : \mu = 0$ and $H_a : \mu = 2$. The chart includes a conventional 2.5% rejection region, on the right tail of the null distribution. **Is there any better way to "spend" a small rejection region, relative to the alternative hypothesis?**
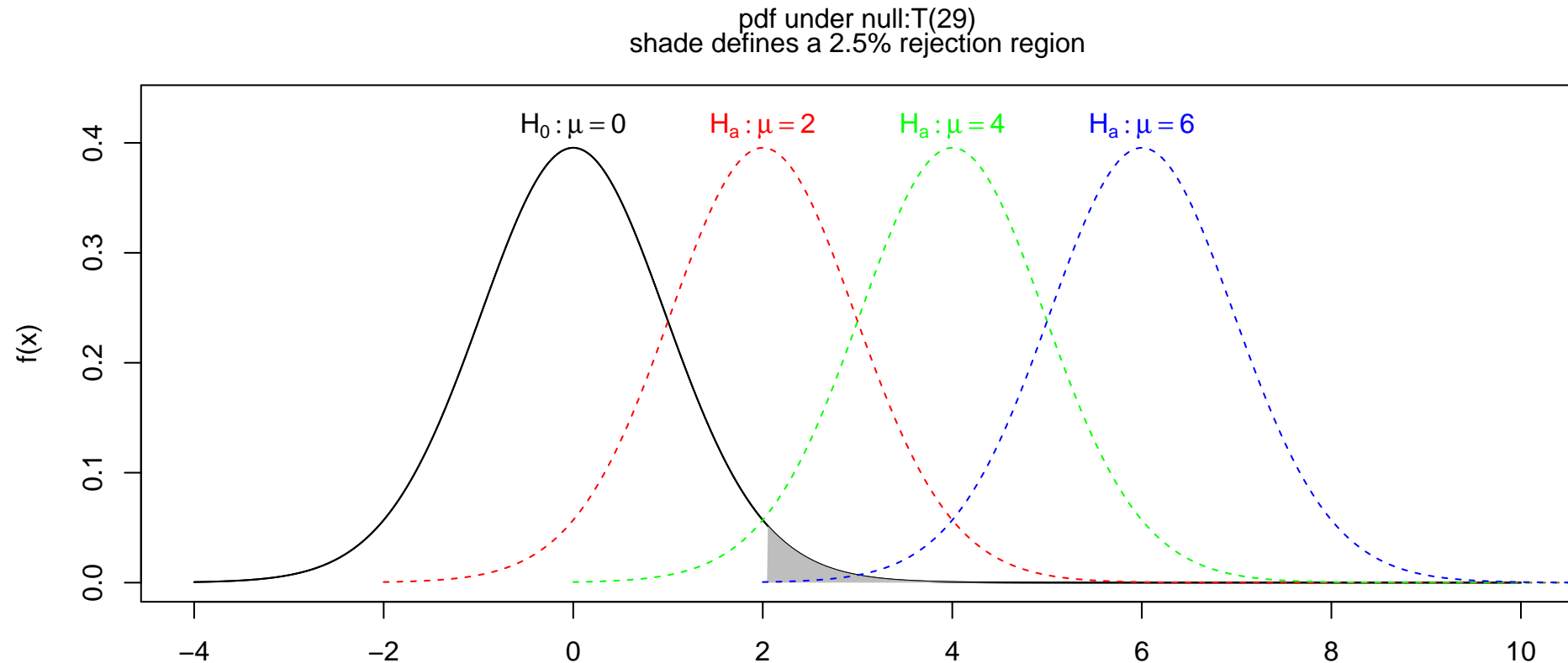
Figure 3.6: A stack of simple alternatives to the null, which is t-distributed.

With reference to figure 3.6, let $H_0 : \mu = 0$ and $H_a : \mu \in \{2, 4, 6\}$ (a composite of the three distinct simple alternatives displayed). The chart includes a conventional 2.5% rejection region, on the right tail of the null distribution. Is there any better way to "spend" a small rejection region, relative to the alternative hypothesis?

Now assume that "Accept H_0" is the incumbent "decision" that you apply to the null hypothesis. Let the t-test be an honest effort to build evidence *against* this position.

Now assume that if the test statistic is extreme enough in its cdf under the null, then your decision is to "Reject H_0".

If you decide to accept or reject a null hypothesis then you leave open the possibility of error, relative to the truth. These errors are conventionally called "Type I" error and "Type II" error. The following table identifies these errors for the context in which only H_0 and H_a are considered.

	H_0 **is True**	H_a **is True**
Accept H_0	✓	Type II error
Reject H_0	Type I error	✓

Below is a list of jargon we have used so far, most of which was introduced in section 3.1.

> **Discussion: Are you clear about the jargon?**
>
> 1. Null hypothesis.
> 2. Alternative hypothesis.
> 3. Model and assumptions that go with it.
> 4. Test statistic and its distribution (& "correct specification").
> 5. Cumulative Density Function.
> 6. Extreme regions of a *cdf*.
> 7. Simple alternative hypotheses (cf. composite).
> 8. Decision processes — accepting and rejecting the null.
> 9. Size and power ... next page.

- **Size:** let the "size" of a test be the probability that an extreme test statistic occurs when the null hypothesis is true. In other words, size is $\mathbb{P}(\text{Type I error})$.

- **Power:** let the "power" of a test be the probability of an extreme test statistic when the alternative hypothesis is *true*. The key to power is the test statistic's **sensitivity** to violations of the null. In other words, power is $1 - \mathbb{P}(\text{Type II error})$.

In a formal Neyman-Pearson context (this pair introduced the concept of alternative hypotheses), power can be defined more generally as the probability of test rejection, in which case we would want to maximise power under the alternative and hold power (aka size) low when the null was true.

Questions

1. Why is power relative to *simple* alternative hypotheses?

2. There is trade off between power and size (think about adjusting the *cdf* rejection regions for the test statistic).

3. If you do not reject the null hypothesis is it interesting to know the power of the test?

We have

$$H_0 : \mu = 0 \tag{3.13}$$

$$H_a : \mu \neq 0 \tag{3.14}$$

and

$$t_x = \frac{\hat{\mu} - 0}{\hat{\sigma}/\sqrt{n}} \sim T(n-1), \tag{3.15}$$

under the null hypothesis and its related assumptions. Assume that we reject the null hypothesis if t_x is within the outer 2.5% quantiles of the $T(n-1)$ distribution (see related discussion of p-values in section 3.1.11 on page 55).

Questions

1. What is the size of this test?

2. What would increase the power of the test, holding size fixed?

The so-called Neyman-Pearson approach to hypothesis testing, with explicit alternatives and fully defined models, is formally presented in Casella and Berger (2002). A more terse formal presentation is found in Wasserman (2010). Finally, the definitive text is Lehmann and Romano (2006).

In these books you will find discussions of power functions, test decision making with explicit loss functions, the Neyman-Pearson lemma and "most powerful tests".

We will not get bogged down in the mathematical statistics foundations of decision theory or Neyman-Pearson's logic (as good as it is). We will spend our time looking at real data and analysing practical limitations of the assumptions behind common tests.

> **Questions**
> 1. Consider again Fisher's set-up without an explicit alternative. What would entail similar inferences to the Neyman-Pearson set-up, without the extra mathematics?

For a counterpoint to the approach of Neyman-Pearson see part three of Fisher (1959).

3.1 Stating The Problem

We now begin the study of the statistical problem that forms the principal subject of this book, the problem of hypothesis testing. As the term suggests, one wishes to decide whether or not some hypothesis that has been formulated is correct. The choice here lies between only two decisions: accepting or rejecting the hypothesis. A decision procedure for such a problem is called a *test* of the hypothesis in question.

The decision is to be based on the value of a certain random variable X, the distribution P_θ of which is known to belong to a class $\mathcal{P} = \{P_\theta, \theta \in \Omega\}$. We shall assume that if θ were known, one would also know whether or not the hypothesis is true. The distributions of $\mathcal{P}$ can then be classified into those for which the hypothesis is true and those for which it is false. The resulting two mutually exclusive classes are denoted by H and K, and the corresponding subsets of Ω by Ω_H and Ω_K respectively, so that $H \cup K = \mathcal{P}$ and $\Omega_H \cup \Omega_K = \Omega$. Mathematically, the hypothesis is equivalent to the statement that P_θ is an element of H. It is therefore convenient to identify the hypothesis with this statement and to use the letter H also to denote the hypothesis. Analogously we call the distributions in K the alternatives to H, so that K is the *class of alternatives*.

Let the decisions of accepting or rejecting H be denoted by d_0 and d_1 respectively. A nonrandomized test procedure assigns to each possible value x of X one of these two decisions and thereby divides the sample space into two complementary regions S_0 and S_1. If X falls into S_0, the hypothesis is accepted; otherwise it is rejected. The set S_0 is called the region of acceptance, and the set S_1 the region of rejection or *critical* region.

Figure 3.7: Extract from Lehmann and Romano (2006), page 56.

Questions

1. If you master the intuition of setting-up and running a simple t-test, are you set to explore the key tests in statistics — including judging regression parameters, assessing prediction models, non-parametric set-ups like the Wilcoxon Rank-Sum test, and ANOVA? [hint: see section 3.1.13 on page 58].

2. How does our test set-up relate to a Bayesian perspective? [hint: see section 3.1.6 on page 50]

3. For cartoon on the next page, can you write down the hypothesis test details (hypotheses, size and power)? And does it matter how often you ask the box to test the hypothesis?

Consider a summary of this section and, within this context, contemplate the following cartoon.

xkcd.com
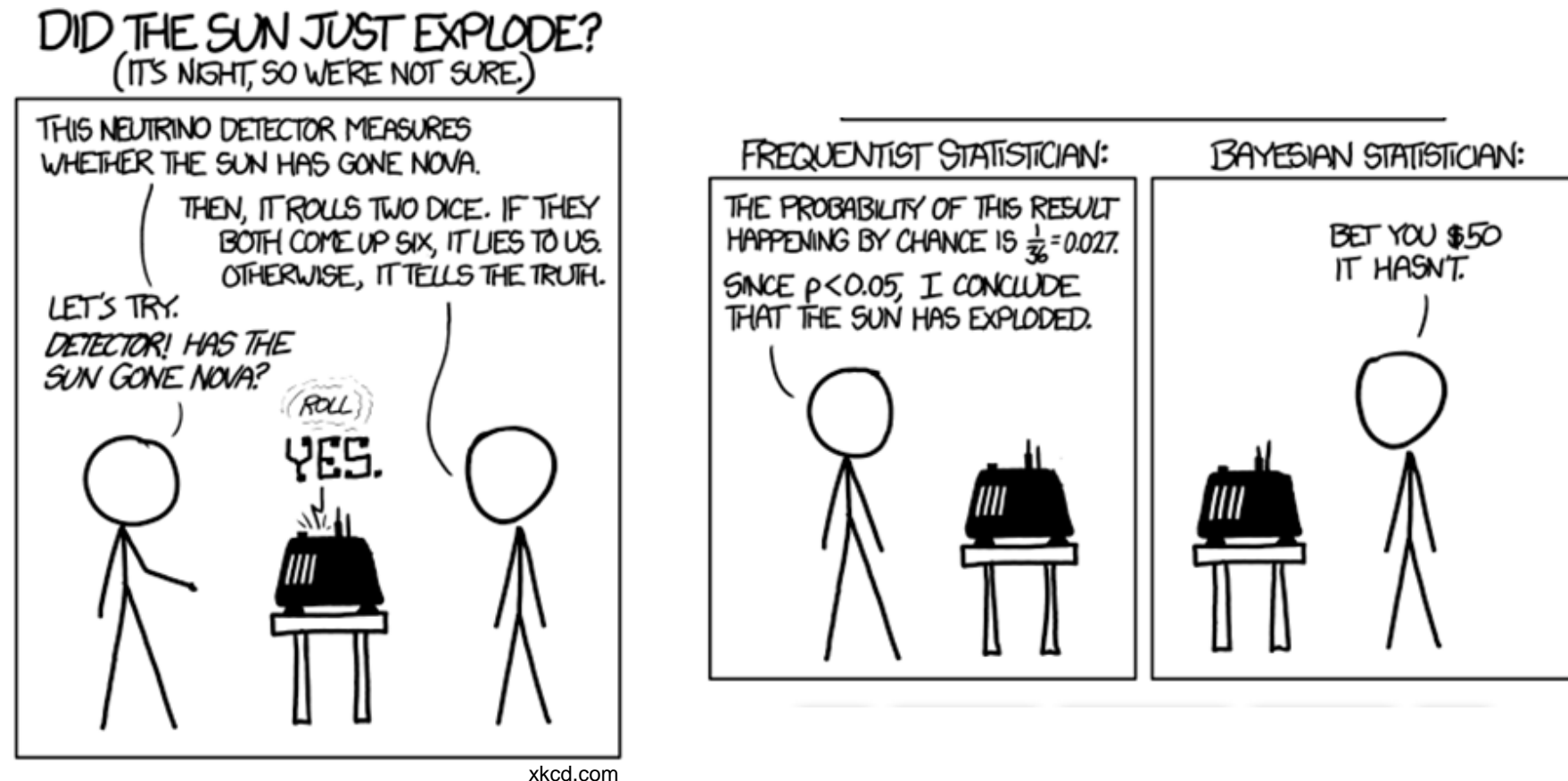
Chapter 4: Appendices

§4.1: Model Assessment

The process of model assessment, comparison and gauging the goodness-of-fit of any particular model is woolly. In this section we try to clarify what the key problems are and how they can be addressed.

This section introduces modern methods that can help us get a handle on how well our models really work. Of particular concern will be the following problems.

- **Over-fitting**: when a model fits in-sample data very well but fails to properly predict out-of-sample data. There is an introduction to over-fitting in section 2.2 on page 38. Here we will split over-fitting in two.

    1. Let *over-fitting*$_1$ represent over-fitting due to a model's **structural flexibility**, in terms of its tendency to "chase" purely random aspects of the in-sample data.

    2. Let *over-fitting*$_2$ represent over-fitting with respect to the **search** processes across potential model forms and data sets.

- **Multiplicity**: when testing many hypotheses at once causes inferential confusion because you know that there will be many extreme-looking p_0-values by chance alone. This problem is logically benign: careful interpretation of p_0-values resolves the confusion.

Multiplicity is intimately linked with *over-fitting*$_2$, since it is useful to produce p_0-values when searching across many models and data sets. The crux of multiplicity is that the true usefulness of p_0-values is laid bare: there is much more to model assessment than p_0-values — not least, prior beliefs and background knowledge.

An example is perhaps the best way to introduce the problem of multiplicity. Assume that you run one hundred separate linear regressions for y against one hundred different x variables. Assume that all of the x-variables are unrelated to y and that each x-variable is independent of the other x-variables. So you would estimate

$$\boldsymbol{y} = \alpha_i + \beta_i \boldsymbol{x_i} + \boldsymbol{\epsilon} \tag{4.1}$$

for $i = 1$ to 100.

Then assume that you calculate a p_0-value for each estimated intercept $\hat{\alpha}_i$, each of which should be akin to a draw from a U(0,1) distribution. Finally, assume that you regard any particular p_0-value, p_i, to be extreme if $p_i < .025$ or $p_i > 0.975$.

Now, the chance of finding *at least one* extreme p_0-value is one minus the chance of finding none. Since the x variables are unrelated to one another, this is $1 - .95^{100} = .994$. Such a high chance of finding an extreme p_0-value (or, to a diminishing extent, many of them) is the essence of the problem of multiplicity.

The classical way to deal with multiplicity is to make p_0-value cut-offs more extreme. For example, in the case with n hypothesis tests the "**Bonferroni**" adjustment is to transform the single case cut-offs of $[.025, .975]$ to $[.025/n, 1 - .025/n]$. This is equivalent to penalising the p_0-value values themselves and leaving the cut-offs fixed.

Appropriately adjusting the cut-offs or p_0-values to account for the fact many hypotheses are being tested simultaneously ensures that an "expected error rate" of extreme result designations is controlled in a meaningful sense.

But this control typically comes at the expense of power — the ability to correctly "identify" violations of null hypotheses is reduced, sometimes dramatically. We will examine several modern methods that attempt to assuage the problems of power and maintain a pre-specified level of error.

But we should keep in mind that a p_0-value alone is never decisive: it is simply a **careful deduction** about the probability of a hypothesis **conditional on the null being true.** From this perspective there is no real need to adjust raw p_0-values when considering many of them — each raw p_0-value is *only part of* an "evidential dossier" on which we *may* base decisions. In other words, p_0-values are summary statistics.

A Bayesian deals directly in the unconditional probability of hypotheses rather than p_0-values (see section 3.1.6 on page 50) — this goes well beyond the p_0-value artifice which is "conditional on the null being true". The Bayesian wants to know the probability of the null hypothesis, *tout court*.

Accordingly, some Bayesians have said that the problem with multiplicity "doesn't really exist" (Lindley, 1997) or that it "just doesn't come up" (Gelman, 2011). But to make useful inferences from large numbers of hypotheses a Bayesian needs an all-singing-all-dancing probability distribution over the entire hypothesis space. Is this always feasible? Where do the priors come from? Will one probability distribution clearly be better than the rest? I don't think Bayesianism yet offers a methodological panacea for the problem of multiplicity — especially when the hypotheses relate to vast model searches.

Not all Bayesians think the same as Lindley and Gelman. In fact, the number one "open problem" in Bayesian statistics, according to the "top 50" Bayesian statisticians, is "[m]odel selection and hypothesis testing"; with Jim Berger going as far as saying "science is choking on the multiplicity problem" (Jordon, 2011). And Gelman later recognised the problem does come up in his experience (Gelman and Loken, 2016) ...

There is no quick-fix for over-fitting.

- Coping with *over-fitting*$_1$ entails measuring how much true randomness in the in-sample data has been erroneously accounted for by an estimated model. In practice we do not know the truth about inherent randomness, this is something we have to estimate (often with the help of ultimately un-testable assumptions). For example, the *true* variance of an asset's returns is unknowable (but we may be able to make a reasonable estimate).

- A sufficient resolution to *over-fitting*$_2$ would be for your beliefs across models (and appropriate data selections) to be calibrated to reality. But thinking that you are so calibrated can be a chimera. For example, assume that you performed a vast data-mining exercise, trying out many different linear and non-linear models, along with different x-variables, in an effort the predict the return on the S&P 500 index. Theoretically, there is no problem with using your model results to work out "optimal" trading decisions, in terms of Subjective Expected Utility Theory (which requires Bayesian-style probability beliefs across the models). In practice, if your beliefs about which models are best are not close to reality, your investment decisions will be bad. But how can you measure how well calibrated your beliefs are?

Below is a menu of modern methods that can help you tackle the confusion of multiplicity, measure over-fitting and calibrate your beliefs about models to reality. But there is no silver-bullet or recipe to follow.

Topic	Details
Out-of-sample testing	In theory, deals with *over-fitting*$_1$, *over-fitting*$_2$ and multiplicity
Prediction Optimism	Designed to deal with *over-fitting*$_1$ and possibly *over-fitting*$_2$
Cross-validation	Designed to deal with *over-fitting*$_1$ and possibly *over-fitting*$_2$
Information Criteria	Designed to deal with *over-fitting*$_1$
False Discovery Rates	Designed to deal with multiplicity and *over-fitting*$_2$
Reality-check p-values	Designed to deal with multiplicity and *over-fitting*$_2$
Bootstrapping	General procedure to help calibrate beliefs
Scenario analysis	General concept related to creativity and beliefs in models

"Ignorance: He that judges without informing himself to the utmost that he is capable, cannot acquit himself of judging amiss."

(Locke, 1690, chapter XXI, paragraph 69)

We will take two things from Locke's quote.

1. Searching across models and data sets is a fundamental requirement of being rational.

2. We need to go beyond statistical hypothesis tests and p-values — logical steps include examining how results change across scenarios and simulation studies.

Exercise 5 (Over-fitting and Multiplicity).

    A. *Revisit the details of p_0-values and hypothesis testing in section 3.1 on page 44. State in your own words what a p-value and p_0-value means.*

    B. *If you understand what a singular p-value really means, what can go wrong when considering many p-values at the same time?*

    C. *Describe why, in general, over-fitting$_1$ is easier to deal with than over-fitting$_2$.*

§4.2: False Discovery Rates

Section Goals ...

- Raise awareness of data-mining dangers and their relation to multiplicity.

- Define the Bonferroni p-value adjustment.

- Explain why FDR helps with multiplicity but is not a panacea.

- Detail the FDR algorithm.

This section briefly introduces data-mining, which is unfortunately a nebulous concept. However, there is an implied link with multiplicity and multiple hypothesis tests. In summary, if you do a large amount of searching you need to adjust your criterion for, or interpretation of, "extreme" p_0-values.

Don't make a fool of yourself

"You make a fool of yourself if you declare that you have discovered something, when all you are observing is random chance. From this point of view, what matters is the probability that, when you find that a result is 'statistically significant', there is actually a real effect. If you find a 'significant' result when there is nothing but chance at play, your result is a false positive, and the chance of getting a false positive is often alarmingly high."

(Colquhoun, 2014, page 1)

In science there is an incentive to publish "significant" results, which arguably motivates data-mining. General examples include the following.

- Searching for rules, patterns or features that group data into apparently-meaningful sub-sets.

- Searching over hundreds of x variables and time series sub-sets for regressions.

- Searching over many different functional forms of models.

The idea behind Bonferroni's adjustment is to control the *expected* error rate of "at least one falsely identified null hypothesis rejection".

In practice, this amounts to straightforwardly adjusting p-values by the number of hypotheses being simultaneously investigated. For example:

- Assume that you would classify a singular p-value as being "extremely low" if it is lower than a cut-off of .05.

- If you then intend to run 100 tests together then the Bonferroni adjusted cut-off is $.05/100 = .0005$.

The Bonferroni adjustment is very conservative — and far too severe when the number of plausible hypotheses is high.

- We would usually be willing to live with a few false discoveries if this increased the chance of making some real discoveries. This is where FDR comes in.

FDR methods, initially introduced by Benjamini and Hochberg (1995), tentatively classify hypothesis test results as "discoveries" if the "test statistics" involved are more extreme than designated cut-offs.

The maximum expected error rate, or *expected proportion* of false discoveries, is "controlled" (in expectation) by adjusting the cut-offs.

Relative to the Bonferroni adjustment, FDR allows for a certain proportion of errors, in order to increase the chance of interesting discoveries. This is appealing.

	Test Statistic inside cut-offs	Test Statistic outside cut-offs	Total
True null hypotheses	U	V	M_0
False null hypotheses	T	S	M_1
Total	$M - R$	R	M

Table 4.1: Each cell in the table represents "counts" based on test statistics from M "well specified" hypothesis tests. Cut-off values, in terms of the test statistic value, are used to find R. The actual number of true and false hypotheses are M_0 and M_1 respectively. The actual number of false discoveries is V and the actual false discovery rate is V/R. In reality only M and R are known. The form of test statistic is arbitrary, but for practical purposes is typically based on the likes of z-scores, t-values or associated p-values. The cut-off values (or associated "regions") for the test statistic are also arbitrary, but must be coherent in terms of how the test statistic is supposed to work.

Step 1:   Set α_f and order the p-values from the set of tests.

$$p_1 \leq p_2 \leq \cdots \leq p_M$$

Step 2:   Find i such that

$$\max_i \left[p_i < \alpha_f \frac{i}{M} \right] \text{ where } i = 1 \text{ to } M.$$

Step 3:   Declare p_i and any p-values lower than p_i to represent "discoveries".

Step 4:   Hope that the actual number of false discoveries is no bigger than "expected".

Table 4.2: Steps to control FDR for M hypothesis tests. Each test is assumed to have a p-value, p_i, which is distributed U(0,1) when the null hypothesis is true. M is as per Table 4.1. The α_f is an arbitrary, or conventionally set, expected error bound. Steps 1-3 encapsulate the method proposed by Benjamini and Hochberg (1995).

Problems exist with FDR methods.

- First, the control of the expected value of false discoveries does not control the variance, or other moments, of V/R (actual error rates can be very different to expectations).

- Secondly, dependence between test statistics may vitiate expectation control and can play havoc with the expected variance of error rates.

- Thirdly, standard FDR methods say little about specific hypotheses (the "error rate" applies to all members in the set of discoveries).

Discoveries in the context of FDR methods are negatively affected by testing too many hypotheses together. Referring to Table 4.2, an arbitrary increase in M makes the cut-offs (Steps 2 and 3) more extreme and increases the number of p_0-values (Step 1). This produces an ambiguous effect on overall discovery declarations: having more extreme cut-offs reduces the number discoveries, whilst additional tests increase the odds of observing p_0-values beyond any cut-off.

- If you add true nulls to the mix the effect is not ambiguous: as M_0 grows, more and more true discoveries will be missed for a fixed error rate.

- The effect of adding to M_1 could be positive or negative for the number of true discoveries S – it depends on the joint probability distribution between the added and original p_0-values.

- Finally, replacing the M in Table 4.2 with a reliable estimate of M_0 would result in more discoveries (on average) whilst retaining control over expected error rates. But the estimation and use of M_0 is ad hoc, especially in non-Bayesian settings.

Much of the recent statistical research on FDR focuses on estimating M_0 e.g. Efron (2008). As explained above, this estimate can be used to replace M in the algorithm from Table 4.2. But it can also play a role in explicit Bayesian methods that are similar to FDR.

Estimating M_0 is analogous to estimating an unknown mixture distribution, so many approaches are potentially useful: the method in Barras et al. (2010) is one of many valid approaches.

Unfortunately, the power of FDR methods to detect true null hypotheses is limited – with or without good estimates of M_0. For example, Storey (2004) reports an empirical study which suggests that the benefits of estimating M_0 are only significant when M_0 is very low relative to M (i.e. when there are many true discoveries to be found).

A different suggestion that improves the power of FDR methods is what Efron (2008) calls "enrichment". Basically, this means adding structural knowledge to a set of test statistics, in order to reduce M.

§4.3: A Little Bit of Mathematics

4.3.1 Basics

- **Law of Large Numbers**: As the sample size gets larger, sample statistics get ever closer to the population characteristics.

- **Central Limit Theorem**: Sample statistics computed from means (such as the means, themselves) are approximately normally distributed, (almost) regardless of the parent distribution. This approximation gets closer as the sample size increases.

Large number theorems are about sample averages $\overline{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$.

Theorem 1 (Weak Law of Large Numbers). *If x_t are IID (identically and independently distributed) with finite mean μ then*

$$\overline{x}_T = \frac{x_1 + x_2 +, ..., + x_T}{T} \xrightarrow{p} \mu$$

Chebychev provides a modified Weak Law of Large Numbers in which the difference between a sample average and the true mean tends to zero, even for non IID data.

Theorem 2 (Central Limit Theorem). *If each x_t is a random draw from the same probability distribution with finite mean μ and finite variance σ^2 and $\bar{x}_T = \frac{1}{T}\sum_{t=1}^{T} x_t$ then*

$$\sqrt{T}(\bar{x}_T - \mu) \xrightarrow{d} N(0, \sigma^2).$$

The result is remarkable because it holds (almost) regardless of the form of the *pdf* for x_t. The distribution does not even need to remotely resemble a normal distribution.

Definition 1 (Convergence in Probability). *The random variable x_t is said to "converge in probability" to c as $t \to \infty$, denoted by $x_t \overset{p}{\to} c$ or $\mathrm{plim}(x_t) = c$, if*

$$\lim_{t \to \infty} \mathbb{P}(|x_t - c| > \varepsilon) = 0, \text{ for any } \varepsilon > 0.$$

Definition 2 (Convergence in Distribution). *The random variable x_t with distribution functions $F_t(x)$ is said to "converge in distribution" to a random variable x with distribution function $F(x)$ as $t \to \infty$, denoted by $x_t \overset{d}{\to} x$, if*

$$\lim_{t \to \infty} F_t(x) = F(x).$$

at every continuity point x of $F(x)$.

Loosely speaking, the delta method is an application of a Taylor series expansion for whatever function is used to estimate a parameter in question. Without loss of generality, assume a univariate case with a parameter estimate $\hat{\theta}$ that is derived from some function g of a random statistic $\hat{z}$, that is

$$\hat{\theta} = g(\hat{z}), \tag{4.2}$$

where $\hat{z}$ has true mean μ_z and true variance σ_z^2. A Taylor series expansion of g around μ_z is

$$g(\hat{z}) = g(\mu_z) + g'(\mu_z)(\hat{z} - \mu_z) + \text{remainder}. \tag{4.3}$$

From equation (4.3) we have the classic results that $\mathbb{E}[\hat{\theta}] \approx g(\mu_z)$ and

$$\text{var}(\hat{\theta}) \approx (g'(\mu_z))^2 \sigma_z^2. \tag{4.4}$$

The "delta method" is usually taken to mean plugging in sample estimates of μ_z and σ_z into (4.4) in order to estimate the variance of the parameter estimate $\hat{\theta}$.

§4.4: Probability Distributions

Section Goals ...

- Define key probability distributions.

- Highlight the simple relationships between these distributions.

Think of "analytic" probability distributions as those you can write down on a piece of paper using straight-forward mathematical notation. They are essentially models, in which there are parameters denoted by Greek letters and "data" which is in the form of random variables. We will assume that you know the appropriate definitions and required conditions of the following.

- Probability density functions "*pdf*".

- Cumulative density functions "*cdf*".

In applied statistics, testing model parameters and overall models requires you to be familiar with the details of at least the Normal distribution, the t-distribution, the F-distribution and the χ^2-distribution.

These definitions follow Wasserman (2010).

- The **sample space** Ω is the set of possible outcomes of an "experiment". Points ω within Ω are **outcomes**, realisations or elements. Subsets of Ω are called **events**. For example if you toss a coin twice then $\Omega = \{HH, HT, TH, TT\}$. The event that the first toss is heads is $A = \{HH, HT\}$.

- We assign a real number $\mathbb{P}(A)$ to every event A. This is called the **probability** of A.

The problem is then to link samples spaces with actual data — this is done with the concept of a random variable.

- A **random variable** is a mapping $X : \Omega \to \mathbb{R}$ that assigns a real number $X(\omega)$ to each outcome ω.

- For example, flip a coin ten times. Let $X(\omega)$ be the number of heads in the sample outcome ω. Assume that $\omega = HHTHHTHHTT$, then $X(\omega) = 6$.

Given a random variable X, we define the cumulative distribution function or *cdf* as follows.

- A **cumulative density function** is the function $F_X : \mathbb{R} \to [0, 1]$ defined by

$$F_X(x) = \mathbb{P}(X \le x) \tag{4.5}$$

The *cdf*, sometimes written as simply F, effectively contains all the information about the random variable. We can now define a probability density function or *pdf*.
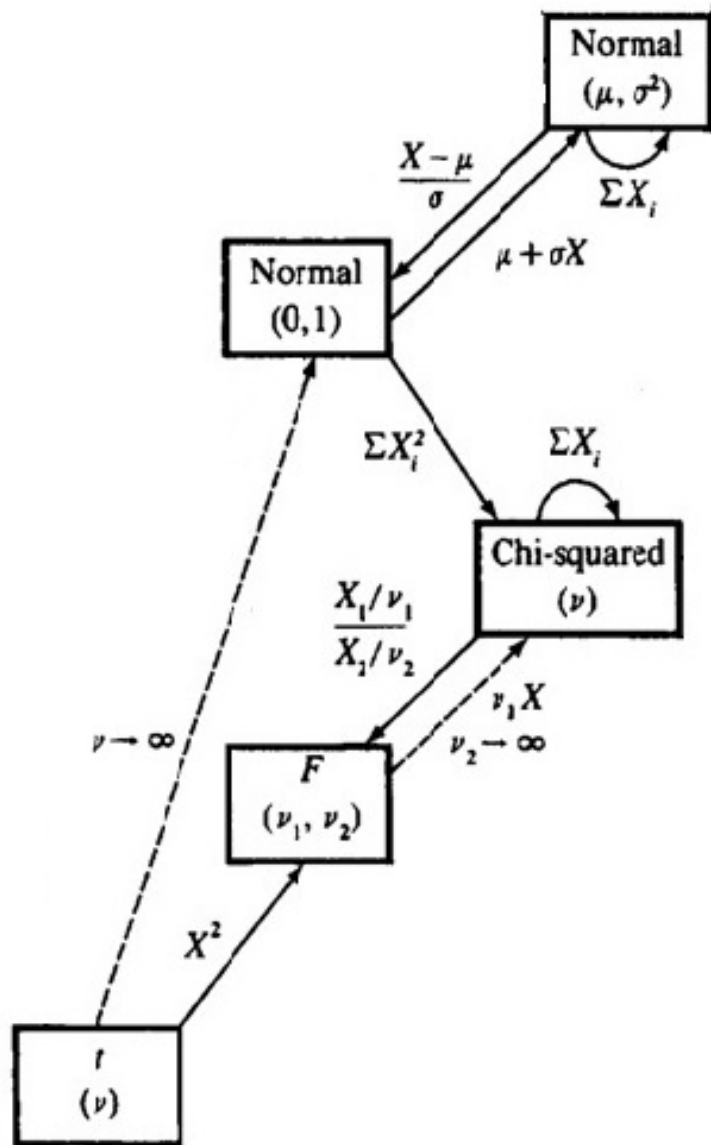
- A random variable X is **continuous** if there exists a function f_X such that $f_X(x) \ge 0$ for all x, $\int_{-\infty}^{\infty} f_X(x)dx = 1$ and for every a and b,

$$\mathbb{P}(a < X < b) = \int_a^b f_X(x)dx. \tag{4.6}$$

The function f_X (or simply f) is called the **probability density function** or *pdf*.

We also have that $F_X(x) = \int_{\infty}^{x} f_X(t)dt$ and $f_X(x) = F_X'$ at all points at which F_X is differentiable.
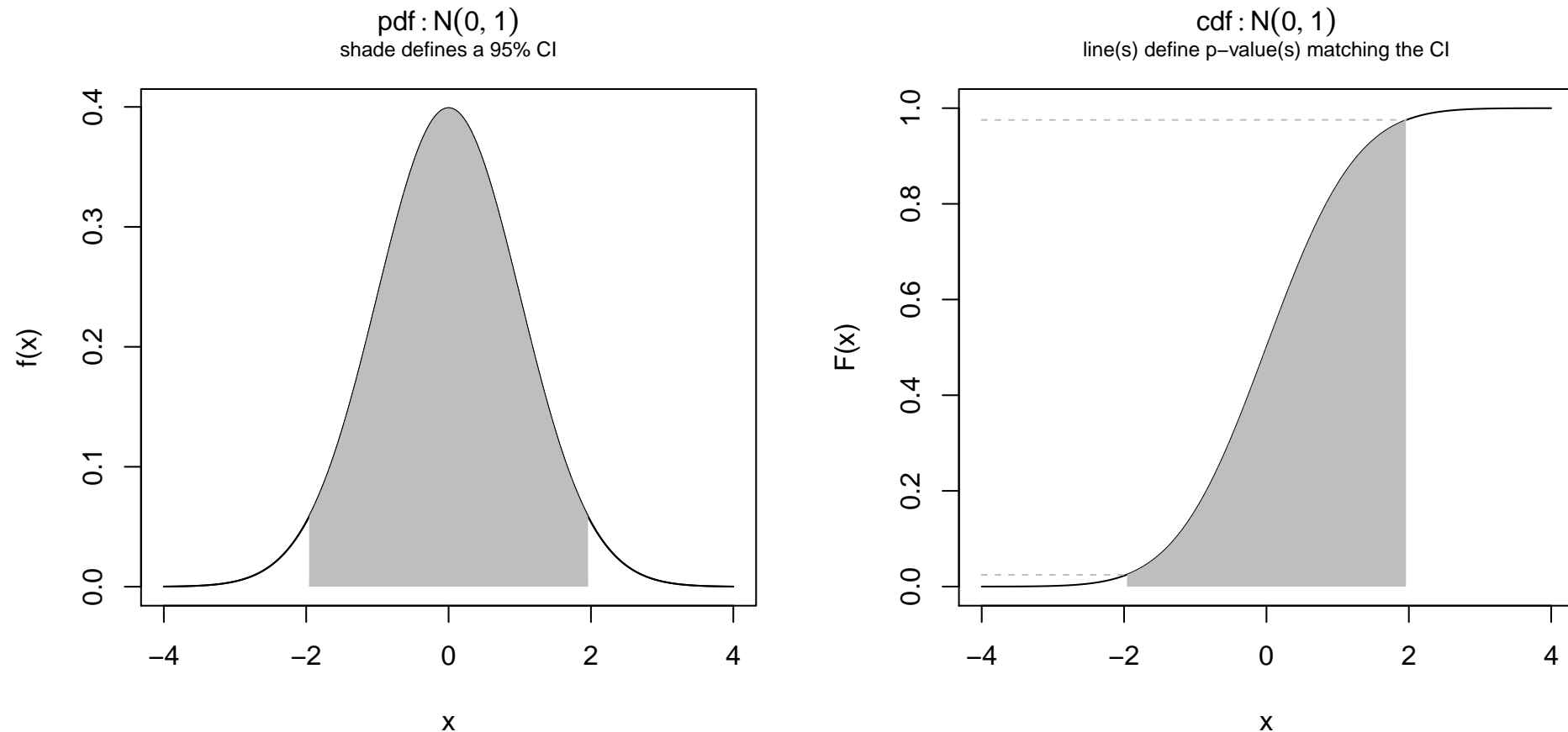
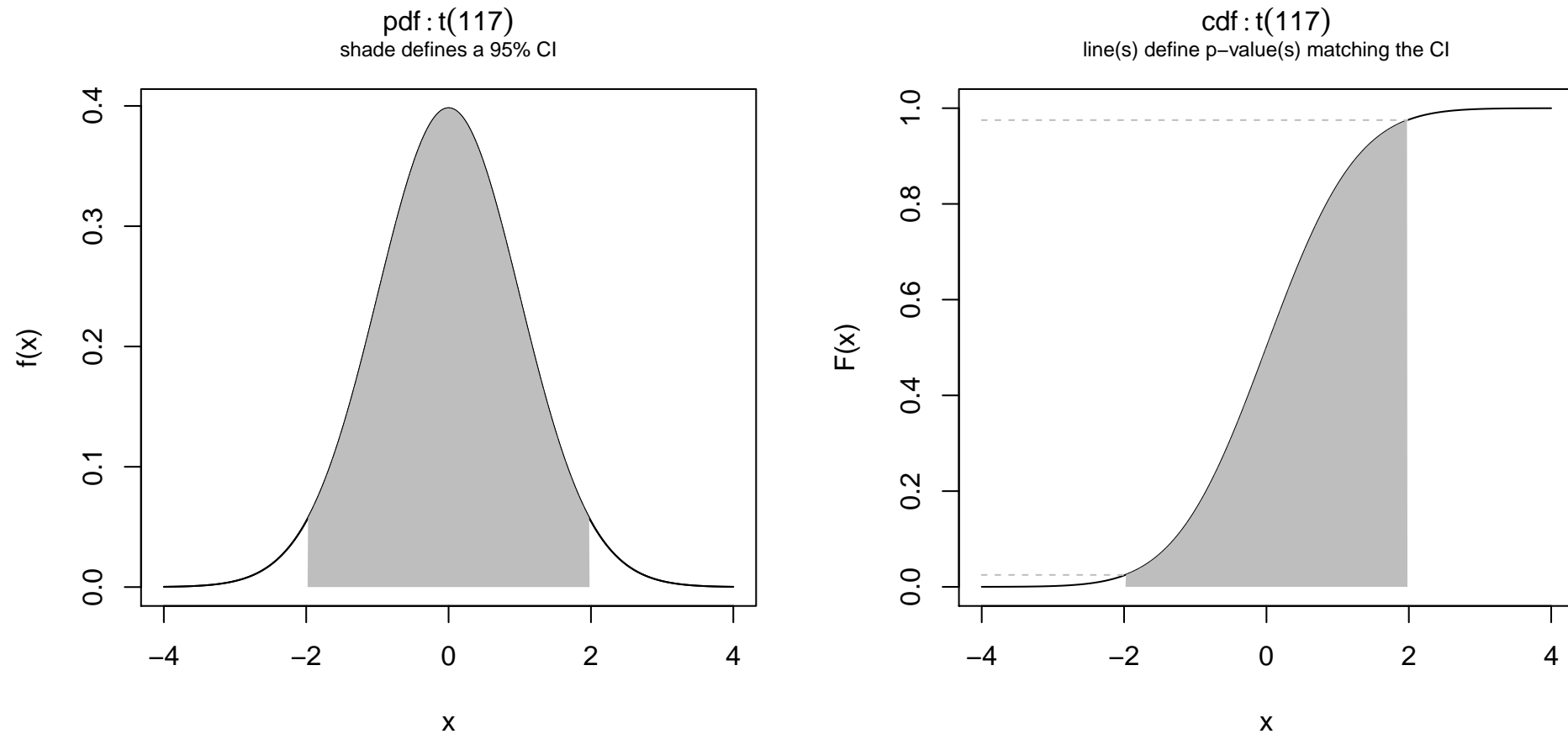The following relationships are useful for statistical testing.

1. If $X \sim N(\mu, \sigma^2)$ then $\frac{X-\mu}{\sigma} \sim N(0,1)$.

2. The $X \sim N(0,1)$ then $X^2 \sim \chi^2(1)$.

3. The sum of K variables that are $\chi^2(1)$ is $\chi^2(K)$.

4. If $X_1 \sim \chi^2(\nu_1)$ and $X_2 \sim \chi^2(\nu_2)$ then $\frac{X_1/\nu_1}{X_2/\nu_2} \sim F(\nu_1, \nu_2)$.

5. If $X \sim F(\nu_1, \nu_2)$ then $\nu_1 X \overset{d}{\to} \chi^2(\nu 1)$ as $\nu_2 \to \infty$.

6. If $X \sim t(\nu)$ then $X \overset{d}{\to} N(0,1)$ as $\nu \to \infty$.

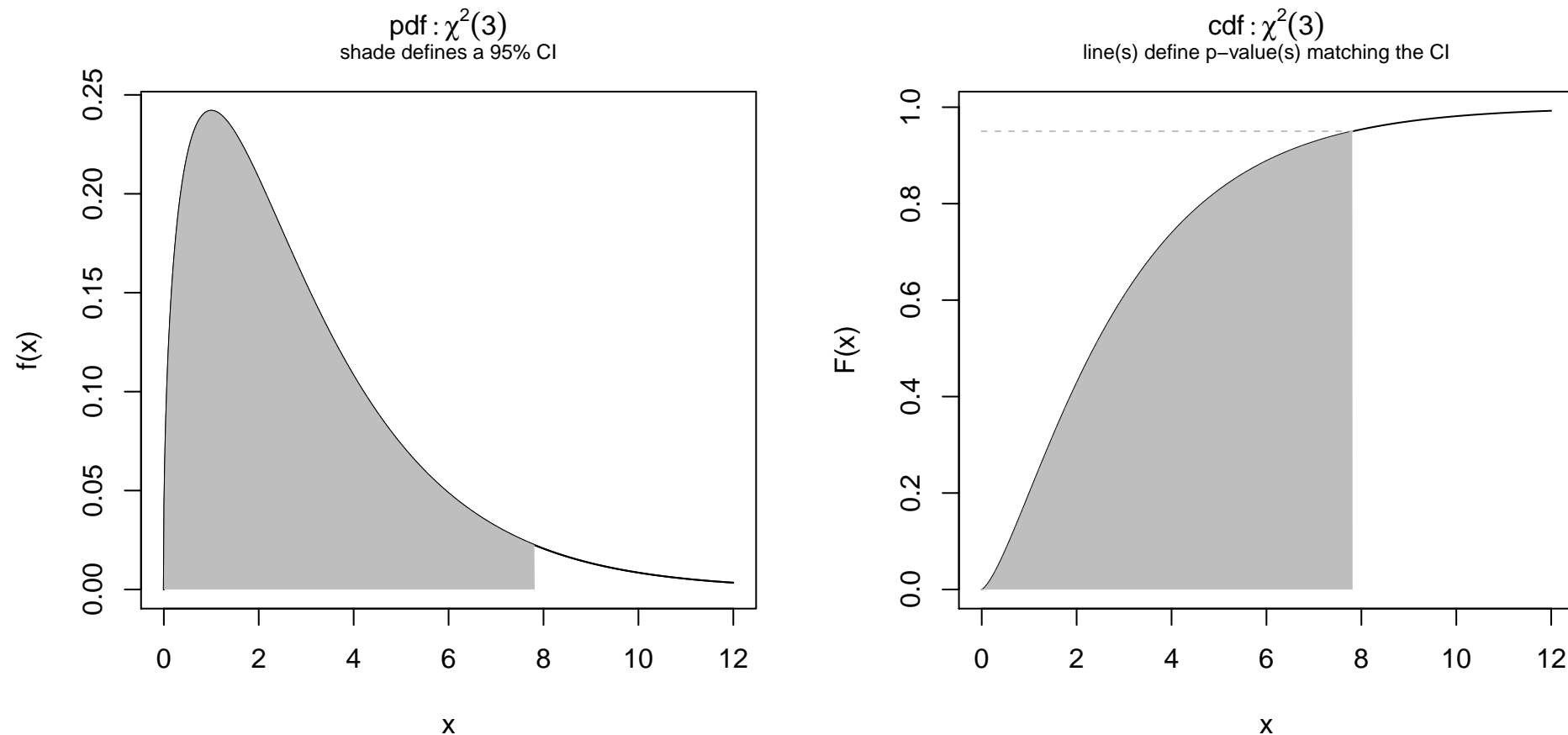7. If $X \sim t(\nu)$ then $X^2 \sim F(1, \nu)$.

4.4.5 PDF Examples

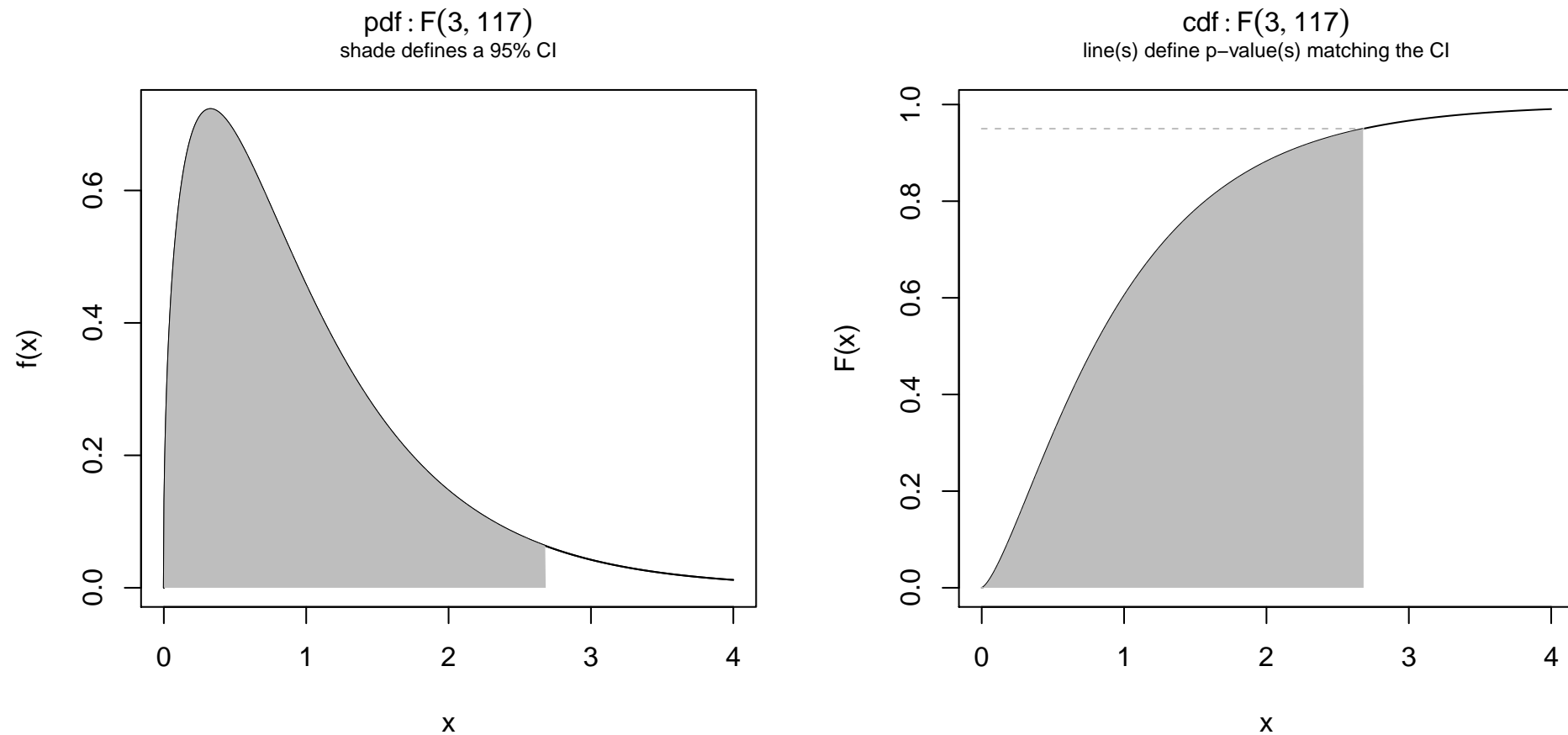**Figure 4.1:** Example of a Normal *pdf* and *cdf*.

Figure 4.2: Example of a t-distribution *pdf* and *cdf*.

Figure 4.3: Example of a χ^2-distribution *pdf* and *cdf*.

4.4.5 PDF Examples

Figure 4.4: Example of an F-distribution *pdf* and *cdf*.

Let an "empirical distribution" be a non-parametric "arrangement" of data that is a counterpart to an analytic distribution. Practical examples for univariate data sets include the following.
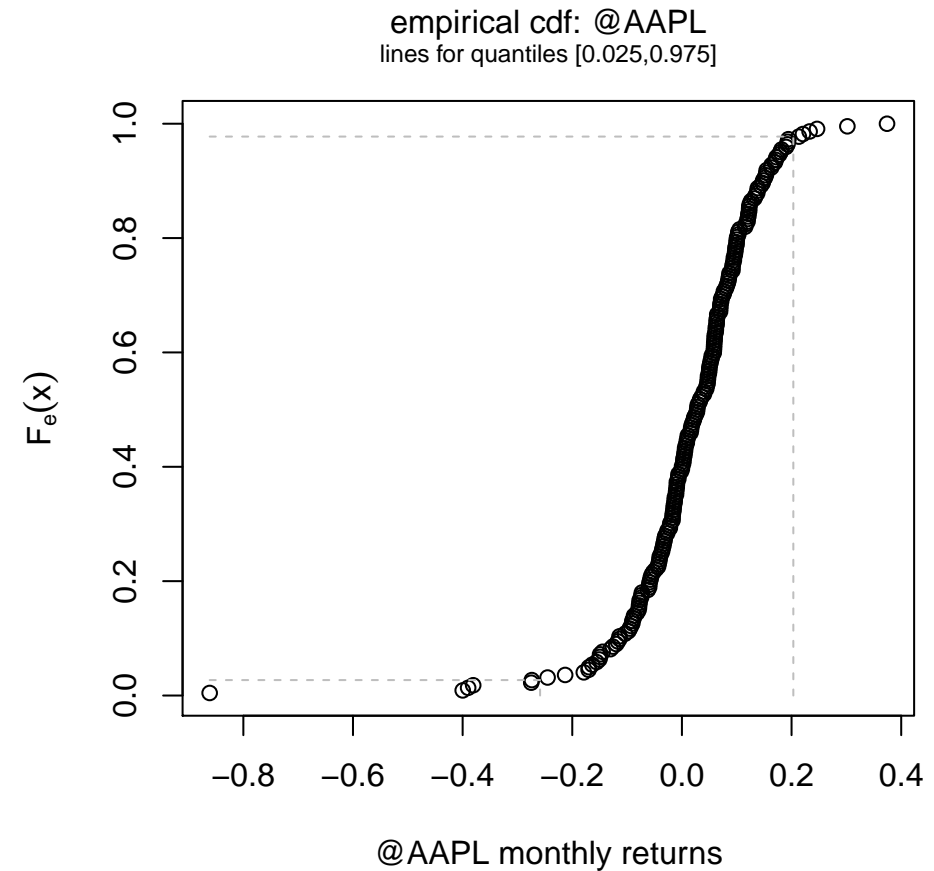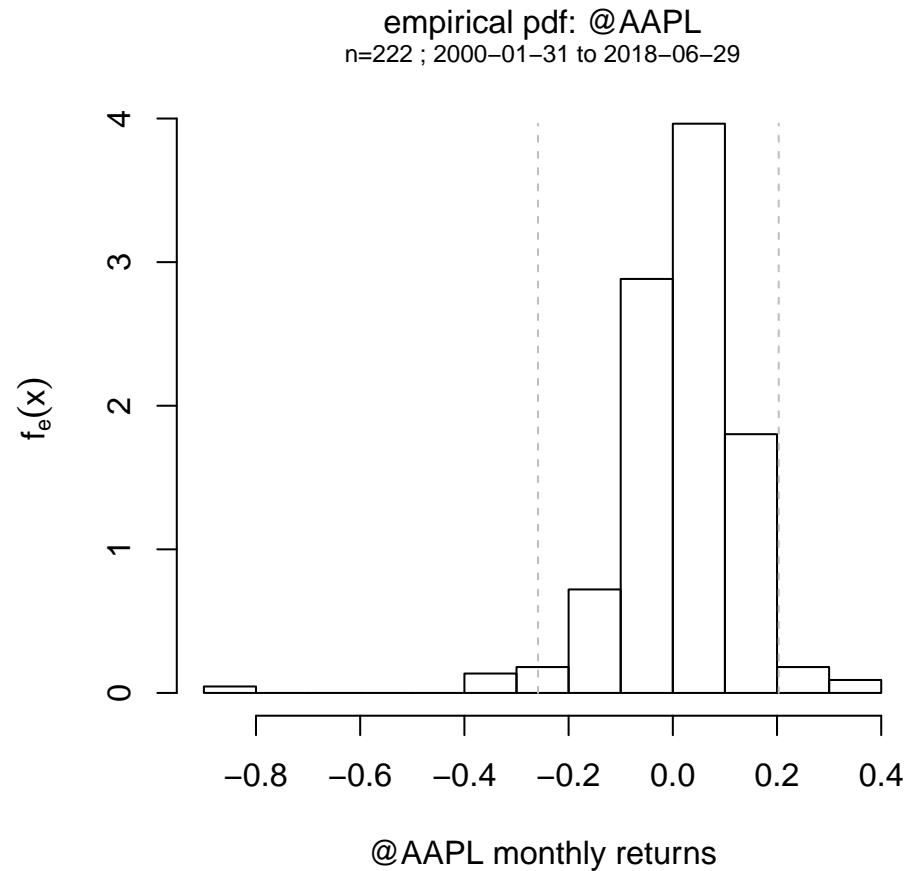
- An empirical *pdf* can be represented by a histogram (scaled to have an "area" of one under its bars).

- An empirical *cdf* is found by ordering the data, which can be conveniently plotted on the x-axis versus the interval [0,1] on the y-axis.

Advanced non-parametric techniques can characterize basic empirical distributions with more detail — though, most of the techniques are essentially approximating "join-the-dots" exercises.

Parametric model fitting can also add detail by replacing Greek letters with numbers. For example, you can assume the data in a histogram is from a Normal distribution and then use Maximum Likelihood Estimation to estimate μ and σ.
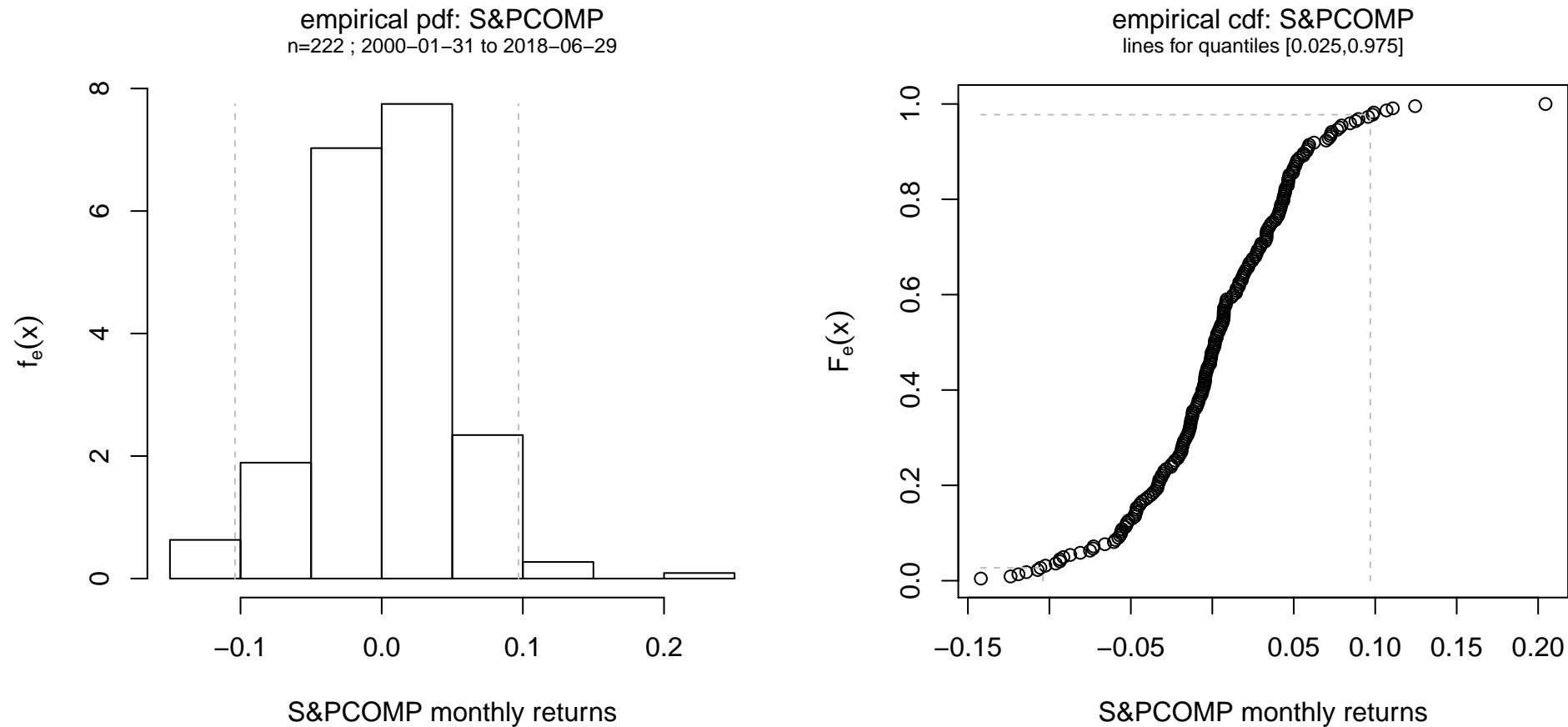
4.4.6 Empirical Distributions

empirical pdf: @AAPL
n=222 ; 2000−01−31 to 2018−06−29

empirical cdf: @AAPL
lines for quantiles [0.025,0.975]

Figure 4.5: Empirical distribution for monthly Apple returns, since 2000.

empirical pdf: S&PCOMP
n=222 ; 2000–01–31 to 2018–06–29

empirical cdf: S&PCOMP
lines for quantiles [0.025,0.975]

Figure 4.6: Empirical distribution for monthly S&P 500 index returns, since 2000.

A quantile-quantile plot ("QQ-plot") is typically used to compare a sample of data with a particular distribution (though two empirical samples could also be compared in the same fashion).

A QQ-plot is simply a scatter plot of empirical quantiles of one dataset versus the quantiles of another. The main trick is to ensure the vectors of numbers are both ordered appropriately. Usually the empirical quantiles are plotted on the y-axis and the comparative distribution quantiles are plotted on the x-axis.

Usually if the two distributions are approximately the same then the QQ-plot shows a relatively straight line. If both the distribution *and* parameters are the same then the line will represent a converging (in sample size) linear relationship with a slope of one and an intercept of zero.

- If the empirical quantiles of y are compared to those implied by a N(0,1) distribution, then look for a line with a slope of var(y) and an intercept of mean(y).

- Conventionally lines are drawn on a QQ-plot by interpolating between the points corresponding to the 25th and 75th quantiles.

Eye-balling a QQ-plot is ad-hoc but useful. And some formal tests on distributional equality are directly related to them (for example the Kolmogorov–Smirnov test).
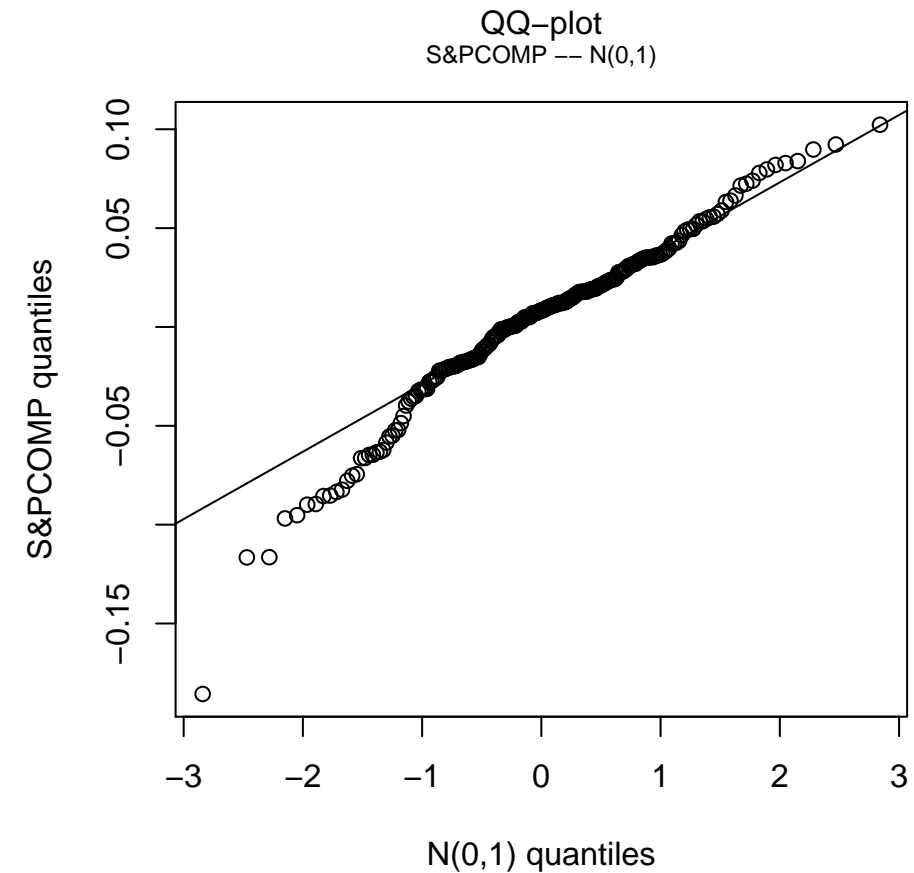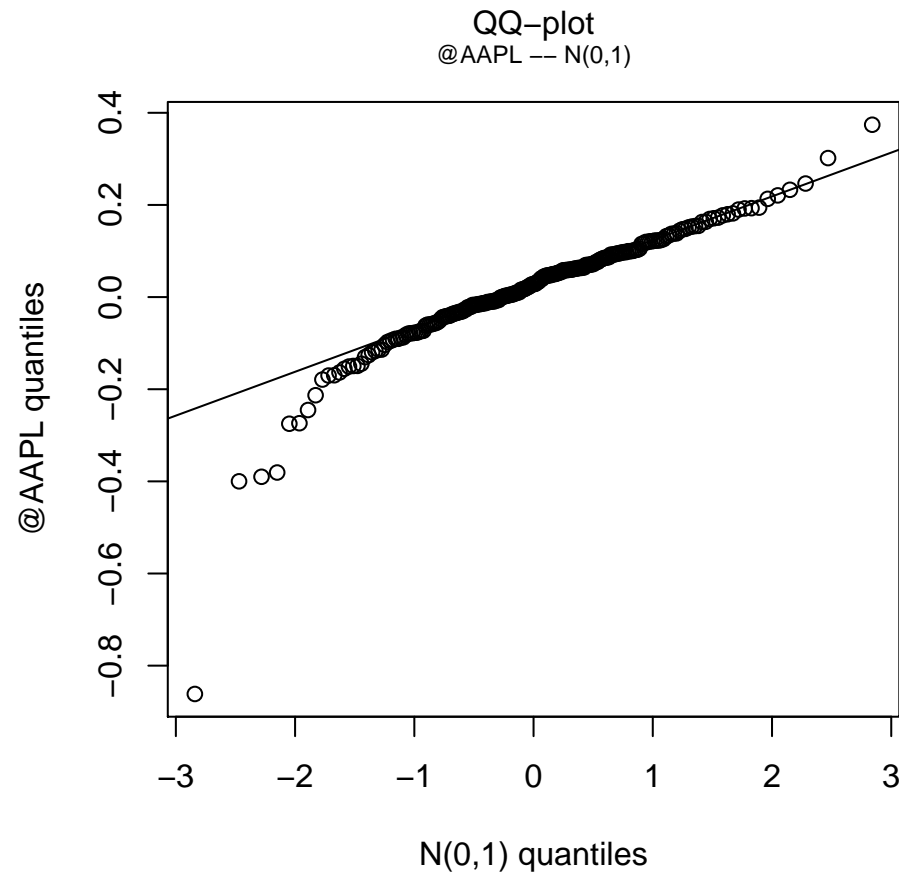
Figure 4.7: QQ Plots - Average monthly returns for Apple and S&P 500 versus N(0,1).

Bibliography

Barras, L., O. Scaillet, and R. Wermers (2010). False discoveries in mutual fund performance: Measuring luck in estimated alphas. *Journal of Finance 65*, 179–216.

Benjamini, Y. and Y. Hochberg (1995). Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society — Series B 57*, 289–300.

Black, F. (1993). Beta and return. *Journal of Portfolio Management*, 75–84.

Casella, G. and R. L. Berger (2002). *Statistical inference* (2 ed.). Duxbury.

Colquhoun, D. (2014). An investigation of the false discovery rate and the misinterpretation of p-values. *Royal Society Open Science 140216*.

Dawid, A. P. (1982). Intersubjective statistical models. In G. Koch and F. Spizzichino (Eds.), *Exchangeability in Probability and Statistics*, pp. 217–232. North-Holland.

Efron, B. (1993). *An Introduction to the Bootstrap*. Chapman & Hall.

Efron, B. and T. Hastie (2016). *Computer age statistical inference*. Cambridge University Press.

Fisher, R., A. (1959). *Statistical Methods, Experimental Design and Scientific Inference* (1990 ed.). Oxford University Press.

Freedman, D., R. Pisani, and R. Purves (2007). *Statistics* (4 ed.). Norton.

Freedman, D. A. (2009). *Statistical models: theory and practice*. Cambridge University Press.

Gelman, A. (2011). Conference presentation. `www.stat.columbia.edu/~martin/Workshop/statistics_neuro_data_931_speaker_04.mov`. Accessed: 2011-06-30.

Gelman, A., J. Carlin, H. Stern, D. Dunson, A. Vehtari, and B. Rubin (2013). *Bayesian Data Analysis* (2 ed.). Chapman & Hall/CRC.

Gelman, A. and E. Loken (2016). The statistical crisis in science. *The best writing on mathematics 2015*, 305.

Hacking, I. (2001). *An Introduction to Probability and Inductive Logic*. Cambridge University Press.

Hansen, L. P. (2013). Nobel prize lecture. `www.nobelprize.org/nobel_prizes/economic-sciences/laureates/2013/hansen-lecture.html`. Accessed: 2017-08-28.

Hastie, T., R. Tibshirani, and J. Friedman (2009). *The elements of statistical learning* (2 ed.). Springer.

Hendry, D. (2011). Empirical economic model discovery and theory evaluation. *RMM - Special Topic: Statistical Science and Philosophy of Science 2*, 115–145.

Hesterberg, T. C. (2015). What teachers should know about the bootstrap: Resampling in the undergraduate statistics curriculum. *The American Statistician 69*(4), 371–386.

James, G., D. Witten, T. Hastie, and R. Tibshirani (2013). *An introduction to statistical learning*. Springer.

Jordon, M. I. (2011). What are the open problems in bayesian statistics? *The International Society for Bayesian Analysis Bulletin 18*, 1–4.

Leamer, E. (1983). Let's take the con out of econometrics. *The American Economic Review 73*, 31–40.

Lehmann, E. and J. Romano (2006). *Testing Statistical Hypotheses*. Springer.

Lindley, D. (1997). Book review of 'multiple comparisons: Theory and methods' by J Hsu. *The Statistician 46*, 572.

Locke, J. (1690). *An Essay Concerning Human Understanding*.

Shalizi, C. (2010). The bootstrap. *American Scientist 98*(3), 186–190.

Siegel, S. and N. J. Castellan (1988). *Nonparametric statistics for the behavioral sciences* (2 ed.). McGraw-Hill.

Spiegelhalter, D. (2019). *The Art of Statistics: Learning from Data*. Penguin UK.

Wasserman, L. (2010). *All of Statistics: A Concise Course in Statistical Inference*. Springer.