Econ 114 Lecture Notes

Advanced Quantitative Methods

Eric M. Aldrich

March 13, 2014

1 Preliminaries **3**
1.1 Probability . . . 3

2 Exploratory Data Analysis **11**
2.1 Kernel Density Estimation . . . 11
2.2 Sample Quantiles . . . 18
2.3 Data Transformations . . . 31

3 Distributions **39**
3.1 Moments . . . 39
3.2 Heavy-Tailed Distributions . . . 46
3.3 Maximum Likelihood Estimation . . . 55

4 Resampling **67**
4.1 Properties of Estimators . . . 67
4.2 Resampling . . . 67
4.3 Resampling . . . 67
4.4 Bootstrapping . . . 67
4.5 Bootstrap Estimates . . . 68
4.6 Bootstrap Estimates . . . 68
4.7 Estimating Bias . . . 68
4.8 Estimating Standard Error . . . 68
4.9 Example: Pareto Distribution . . . 69
4.10 Example: Pareto Distribution . . . 69
4.11 Example: Pareto Distribution . . . 70
4.12 Example: Pareto Distribution . . . 71
4.13 Example: Pareto Distribution . . . 71
4.14 Example: Pareto Distribution . . . 71
4.15 Example: Pareto Distribution . . . 71
4.16 Example: Pareto Distribution . . . 72
4.17 Bootstrap Confidence Intervals . . . 72
4.18 Bootstrap Confidence Intervals . . . 72

5 Time Series **73**
5.1 Stationarity . . . 73
5.2 Autoregressive Processes . . . 77
5.3 Moving Average Processes . . . 87
5.4 ARMA Processes . . . 91

6 Bayesian Methods **93**
6.1 Bayes Theorem . . . 93

6.2 Prior and Posterior Distributions 110

Contents:

CHAPTER

ONE

PRELIMINARIES

Contents:

1.1 Probability

1.1.1 Random Variables

Suppose X is a random variable which can take values $x \in \mathcal{X}$.

- X is a discrete r.v. if $\mathcal{X}$ is countable.
  - $p(x)$ is the probability of a value x and is called the probability mass function.
- X is a continuous r.v. if $\mathcal{X}$ is uncountable.
  - $f(x)$ is called the probability density function and can be thought of as the probability of a value x.

1.1.2 Probability Mass Function

For a discrete random variable the *probability mass function* (PMF) is

$$p(a) = P(X = a),$$

where $a \in \mathbb{R}$.

1.1.3 Probability Density Function

If $B = (a, b)$

$$P(X \in B) = P(a \leq X \leq b) = \int_a^b f(x)dx.$$

Strictly speaking

$$P(X = a) = \int_a^a f(x)dx = 0,$$

but we may (intuitively) think of $f(a) = P(X = a)$.

1.1.4 Properties of Distributions

For discrete random variables

- $p(x) \geq 0, \forall x \in \mathcal{X}$.
- $\sum_{x \in \mathcal{X}} p(x) = 1$.

For continuous random variables

- $f(x) \geq 0, \forall x \in \mathcal{X}$.
- $\int_{x \in \mathcal{X}} f(x)dx = 1$.

1.1.5 Cumulative Distribution Function

For discrete random variables the cumulative distribution function (CDF) is

- $F(a) = P(X \leq a) = \sum_{x \leq a} p(x)$.

For continuous random variables the CDF is

- $F(a) = P(X \leq a) = \int_{-\infty}^{a} f(x)dx$.

1.1.6 Expected Value

For a discrete r.v. X, the expected value is

$$E[X] = \sum_{x \in \mathcal{X}} xp(x).$$

For a continuous r.v. X, the expected value is

$$E[X] = \int_{x \in \mathcal{X}} x\, f(x)dx.$$

1.1.7 Expected Value

If $Y = g(X)$, then

- For discrete r.v. X

$$E[Y] = E[g(X)] = \sum_{x \in \mathcal{X}} g(x)p(x).$$

- For continuous r.v. X

$$E[Y] = E[g(X)] = \int_{x \in \mathcal{X}} g(x)f(x)dx.$$

1.1.8 Properties of Expectation

For random variables X and Y and constants $a, b \in \mathbb{R}$, the expected value has the following properties (for both discrete and continuous r.v.'s):

- $E[aX + b] = aE[X] + b$.

- $E[X+Y] = E[X] + E[Y]$.

Realizations of X, denoted by x, may be larger or smaller than $E[X]$.

- If you observed many realizations of X, $E[X]$ is roughly an average of the values you would observe.

1.1.9 Properties of Expectation - Proof

$$\begin{aligned} E[aX+b] &= \int_{-\infty}^{\infty} (ax+b) f(x) dx \\ &= \int_{-\infty}^{\infty} ax f(x) dx + \int_{-\infty}^{\infty} b f(x) dx \\ &= a \int_{-\infty}^{\infty} x f(x) dx + b \int_{-\infty}^{\infty} f(x) dx \\ &= a\, E[X] + b. \end{aligned}$$

1.1.10 Variance

Generally speaking, variance is defined as

$$Var(X) = E\left[(X - E[X])^2\right].$$

If X is discrete:

$$Var(X) = \sum_{x \in \mathcal{X}} (x - E[X])^2 p(x).$$

If X is continuous:

$$Var(X) = \int_{x \in \mathcal{X}} (x - E[X])^2 f(x) dx$$

1.1.11 Variance

Using the properties of expectations, we can show $Var(X) = E[X^2] - E[X]^2$:

$$\begin{aligned} Var(X) &= E\left[(X - E[X])^2\right] \\ &= E\left[X^2 - 2XE[X] + E[X]^2\right] \\ &= E[X^2] - 2E[X]E[X] + E[X]^2 \\ &= E[X^2] - E[X]^2. \end{aligned}$$

1.1.12 Standard Deviation

The standard deviation is simply

$$Std(X) = \sqrt{Var(X)}.$$

- $Std(X)$ is in the same units as X.
- $Var(X)$ is in units squared.

1.1.13 Covariance

For two random variables X and Y, the covariance is generally defined as

$$Cov(X,Y) = E\left[(X - E[X])(Y - E[Y])\right]$$

Note that $Cov(X,X) = Var(X)$.

1.1.14 Covariance

Using the properties of expectations, we can show

$$Cov(X,Y) = E[XY] - E[X]E[Y].$$

This can be proven in the exact way that we proved

$$Var(X) = E[X^2] - E[X]^2.$$

In fact, note that

$$\begin{aligned} Cov(X,X) &= E[XY] - E[X]E[Y] \\ &= E[X^2] - E[X]^2 = Var(X). \end{aligned}$$

1.1.15 Properties of Variance

Given random variables X and Y and constants $a, b \in \mathbb{R}$,

$$Var(aX + b) = a^2 Var(X).$$

$$Var(aX + bY) = a^2 Var(X) + b^2 Var(Y) + 2abCov(X,Y).$$

The latter property can be generalized to

$$Var\left(\sum_{i=1}^{n} a_i X_i\right) = \sum_{i=1}^{n} a_i^2 Var(X_i) + 2\sum_{i=1}^{n-1}\sum_{j=i+1}^{n} a_i a_j Cov(X_i, X_j).$$

1.1.16 Properties of Variance - Proof

$$\begin{aligned} Var(aX + bY) &= E\left[(aX + bY)^2\right] - E\left[aX + bY\right]^2 \\ &= E[a^2X^2 + b^2Y^2 + 2abXY] - \left(aE[X] + bE[Y]\right)^2 \\ &= a^2E[X^2] + b^2E[Y^2] + 2abE[XY] \\ &\qquad - a^2E[X]^2 - b^2E[Y]^2 - 2abE[X]E[Y] \\ &= a^2\left(E[X^2] - E[X]^2\right) + b^2\left(E[Y^2] - E[Y]^2\right) \\ &\qquad + 2ab\left(E[XY] - E[X]E[Y]\right) \\ &= a^2Var(X) + b^2Var(Y) + 2abCov(X,Y). \end{aligned}$$

1.1.17 Properties of Covariance

Given random variables W, X, Y and Z and constants $a, b \in \mathbb{R}$,

$$Cov(X, a) = 0.$$

$$Cov(aX, bY) = abCov(X, Y).$$

$$Cov(W + X, Y + Z) = Cov(W, Y) + Cov(W, Z) + Cov(X, Y) + Cov(X, Z).$$

The latter two can be generalized to

$$Cov\left(\sum_{i=1}^{n} a_i X_i, \sum_{j=1}^{m} b_j Y_j\right) = \sum_{i=1}^{n}\sum_{j=1}^{m} a_i b_j Cov(X_i, Y_j).$$

1.1.18 Correlation

Correlation is defined as

$$Corr(X, Y) = \frac{Cov(X, Y)}{Std(X)Std(Y)}.$$

- It is fairly easy to show that $-1 \leq Corr(X, Y) \leq 1$.
- The properties of correlations of sums of random variables follow from those of covariance and standard deviations above.

1.1.19 Normal Distribution

The normal distribution is often used to approximate the probability distribution of returns.

- It is a continuous distribution.
- It is symmetric.
- It is fully characterized by μ (mean) and σ (standard deviation) – i.e. if you only tell me μ and σ, I can draw every point in the distribution.

1.1.20 Normal Density

If X is normally distributed with mean μ and standard deviation σ, we write

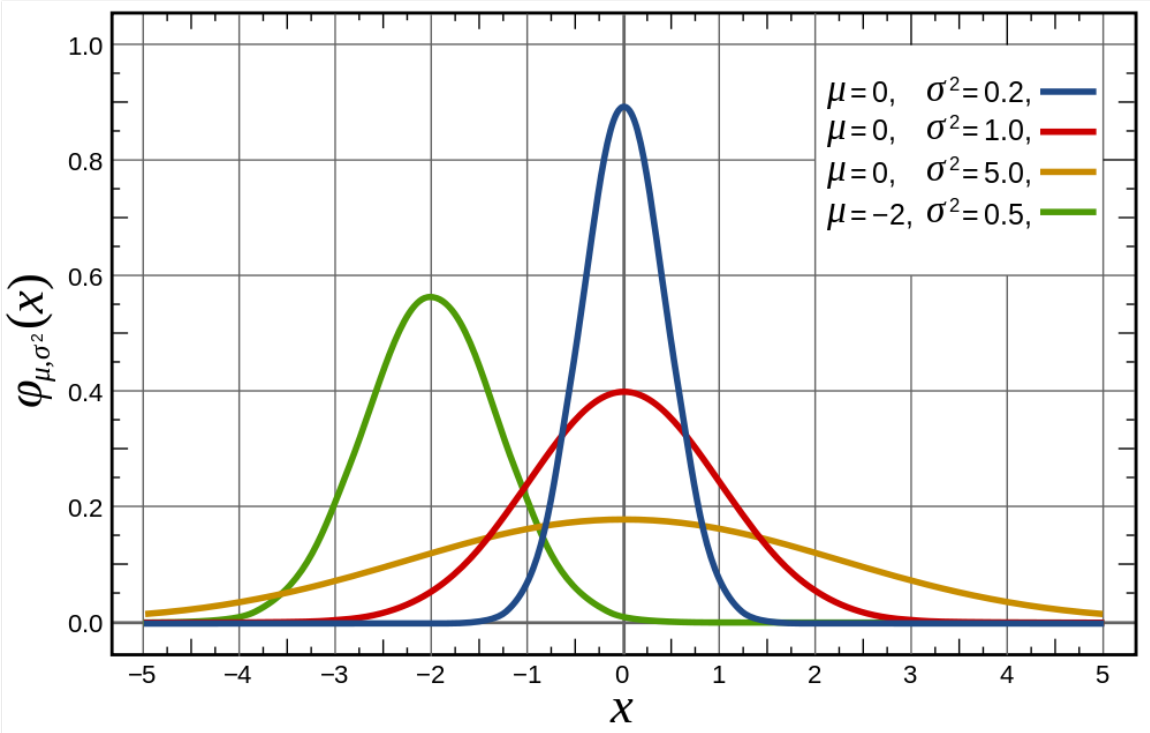
$$X \sim \mathcal{N}(\mu, \sigma).$$

The probability density function is

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

1.1.21 Normal Distribution

From Wikipedia:



1.1.22 Standard Normal Distribution

Suppose $X \sim \mathcal{N}(\mu, \sigma)$.

Then

$$Z = \frac{X - \mu}{\sigma}$$

is a standard normal random variable: $Z \sim \mathcal{N}(0, 1)$.

- That is, Z has zero mean and unit standard deviation.

We can reverse the process by defining

$$X = \mu + \sigma Z.$$

1.1.23 Standard Normal Distribution - Proof

$$\begin{aligned} E[Z] &= E\left[\frac{X-\mu}{\sigma}\right] \\ &= \frac{1}{\sigma}E[X-\mu] \\ &= \frac{1}{\sigma}(E[X]-\mu) \\ &= \frac{1}{\sigma}(\mu-\mu) \\ &= 0. \end{aligned}$$

1.1.24 Standard Normal Distribution - Proof

$$\begin{aligned} Var(Z) &= Var\left(\frac{X-\mu}{\sigma}\right) \\ &= Var\left(\frac{X}{\sigma}-\frac{\mu}{\sigma}\right) \\ &= \frac{1}{\sigma^2}Var(X) \\ &= \frac{\sigma^2}{\sigma^2} \\ &= 1. \end{aligned}$$

1.1.25 Sum of Normals

Suppose $X_i \sim \mathcal{N}(\mu_i, \sigma_i)$ for $i = 1, \ldots, n$.

Then if we denote $W = \sum_{i=1}^{n} X_i$

$$W \sim \mathcal{N}\left(\sum_{i=1}^{n}\mu_i, \sqrt{\sum_{i=1}^{n}\sigma_i^2 + 2\sum_{i=1}^{j}\sum_{j=1}^{n}Cov(X_i, X_j)}\right).$$

How does this simplify if $Cov(X_i, X_j) = 0$ for $i \neq j$?

1.1.26 Sample Mean

Suppose we don't know the true probabilities of a distribution, but would like to estimate the mean.

- Given a sample of observations, $\{x_i\}_{i=1}^{n}$, of random variable X, we can estimate the mean by

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} x_i.$$

- This is just a simple arithmetic average, or a probability weighted average with equal probabilities: $\frac{1}{n}$.
- But the true mean is a weighted average using actual (most likely, unequal) probabilities. How do we reconcile this?

1.1.27 Sample Mean (Cont.)

Given that the sample $\{x_i\}_{i=1}^n$ was drawn from the distribution of X, the observed values are inherently weighted by the true probabilities (for large samples).

- More values in the sample will be drawn from the higher probability regions of the distribution.
- So weighting all of the values equally will naturally give more weight to the higher probability outcomes.

1.1.28 Sample Variance

Similarly, the sample variance can be defined as

$$\hat{\sigma}^2 = \frac{1}{n-1}\sum_{i=1}^{n}(x_i - \hat{\mu})^2.$$

Notice that we use $\frac{1}{n-1}$ instead of $\frac{1}{n}$ for the sample average.

- This is because a simple average using $\frac{1}{n}$ underestimates the variability of the data because it doesn't account for extra error involved in estimating $\hat{\mu}$.

1.1.29 Other Sample Moments

Sample standard deviations, covariances and correlations are computed in a similar fashion.

- Use the definitions above, replacing expectations with simple averages.

CHAPTER

TWO

EXPLORATORY DATA ANALYSIS

Contents:

2.1 Kernel Density Estimation

2.1.1 Data Samples

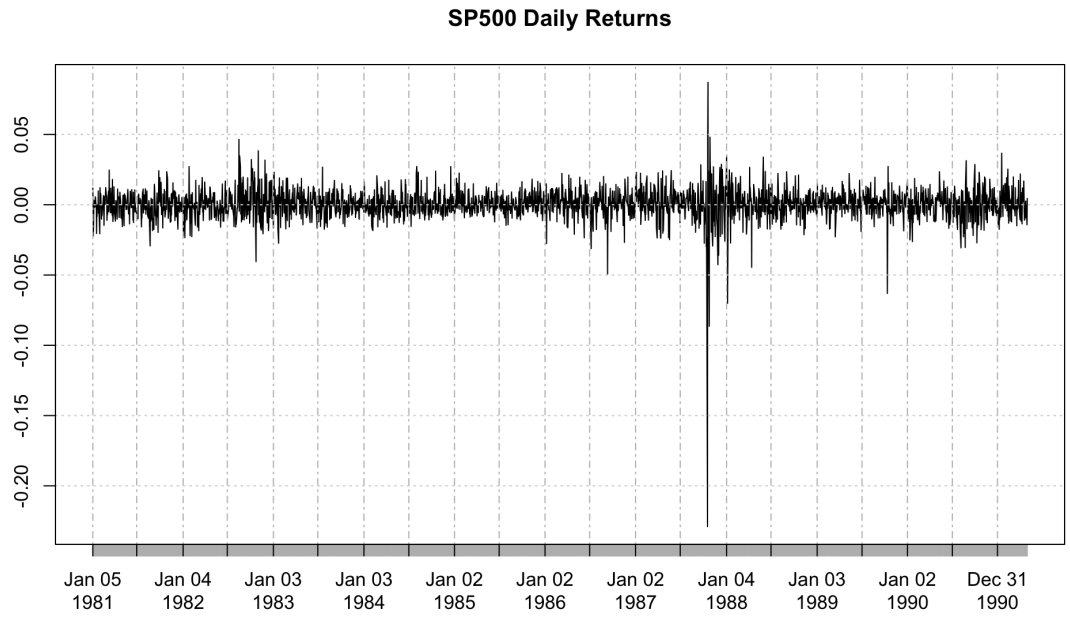
Suppose we are presented with a set of data observations $y_1, y_2, \ldots, y_n$.

- In this course we will often assume that the observations are realizations of random variables $Y_1, Y_2, \ldots, Y_n$, where $Y_i \sim F, \forall i$.
- That is, we will assume Y_i all come from the same distribution.
- In the case of financial data, we will also view the observations $y_1, y_2, \ldots, y_n$ as being ordered by time.
- This is referred to as a *time series*.

The code for generating the plots of this section will be provided below. To run the code, first load data into R with the following script.:

```
# Load the Ecdat package which has the SP500 data used in the text
#install.packages("quantmod")
library(quantmod)

# Get the price data
getSymbols("^GSPC", from="1981-01-01", to="1991-04-30")

# Compute returns from difference of log adjusted closing prices
spRets = diff(log(GSPC$GSPC.Adjusted))
spRets = spRets[-1,]
```

2.1.2 Time Series Data Example



To create this plot, run the following script:

```
# Plot the time series
par(bg="white")
plot(spRets, main="SP500 Daily Returns")
dev.copy(png, file="sp500TimeSeries.png", height=6, width=10, units='in', res=200)
graphics.off()
```

2.1.3 Histogram

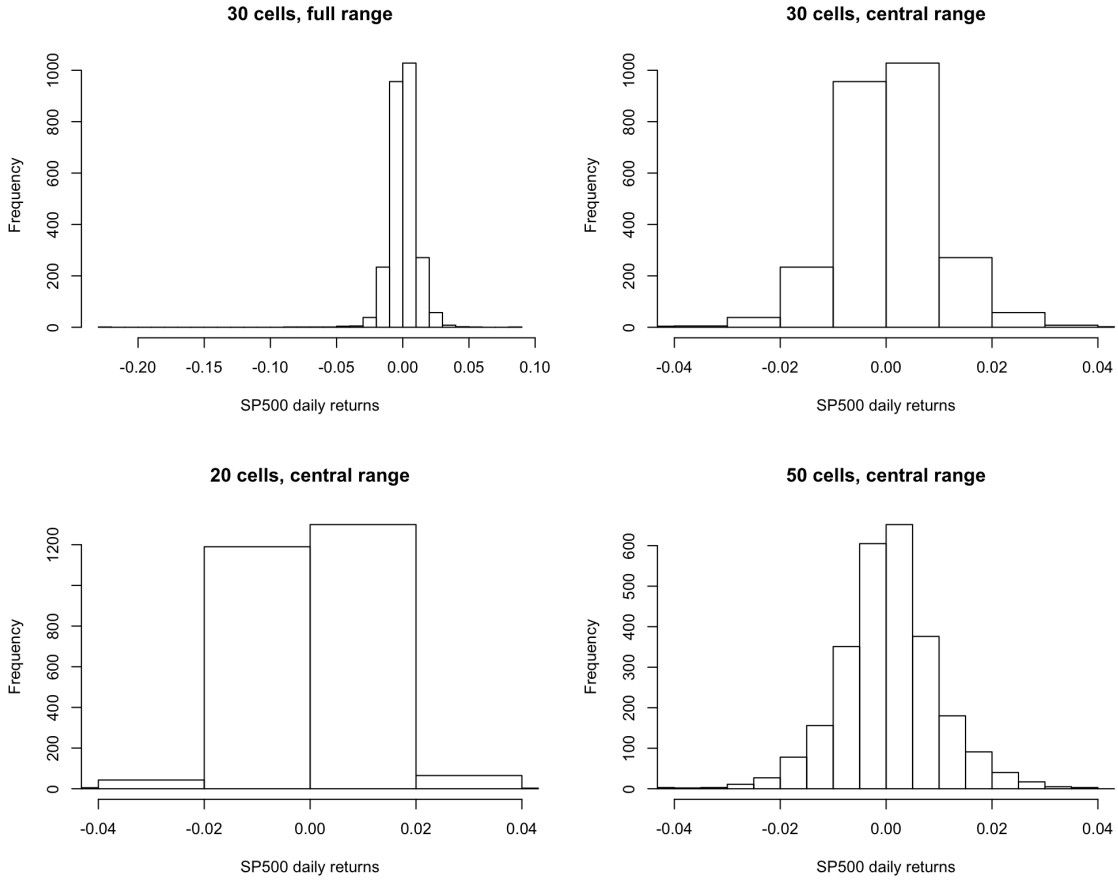
Suppose we have dataset $y_1, y_2, \ldots, y_n$ drawn from the same distribution F.

- We typically don't know F and would like to estimate it with the data.
- A simple estimate of F is obtained by a histogram.

A histogram:

- Divides the possible values of the random variable(s) y into K regions, called bins.
- Counts the number of observations that fall into each bin.

2.1.4 Histogram of SP500 Daily Returns



To create this plot, run the following script:

```
# Plot histograms
par(mfrow=c(2,2), bg="white")
hist(spRets, breaks=30, xlab="SP500 daily returns", main="30 cells, full range")
hist(spRets, breaks=30, xlim=c(-0.04,0.04), xlab="SP500 daily returns",
     main="30 cells, central range")
hist(spRets, breaks=20, xlim=c(-0.04,0.04), xlab="SP500 daily returns",
     main="20 cells, central range")
hist(spRets, breaks=50, xlim=c(-0.04,0.04), xlab="SP500 daily returns",
     main="50 cells, central range")
dev.copy(png, file="sp500Hist.png", height=8, width=10, units='in', res=200)
graphics.off()
```

2.1.5 Kernel Density Estimation

Histograms are crude estimators of density functions.

- The *kernel density estimator* (KDE) is a better estimator.
- A KDE uses a kernel, which is a probability density function symmetric about zero.
- Often, the kernel is chosen to be a standard normal density.

- The kernel has *nothing* to do with the true density of the data (i.e. choosing a normal kernel doesn't mean the data is normally distributed).

2.1.6 Kernel Density Estimation

Given random variables $Y_1, Y_2, \ldots, Y_n$, the KDE is

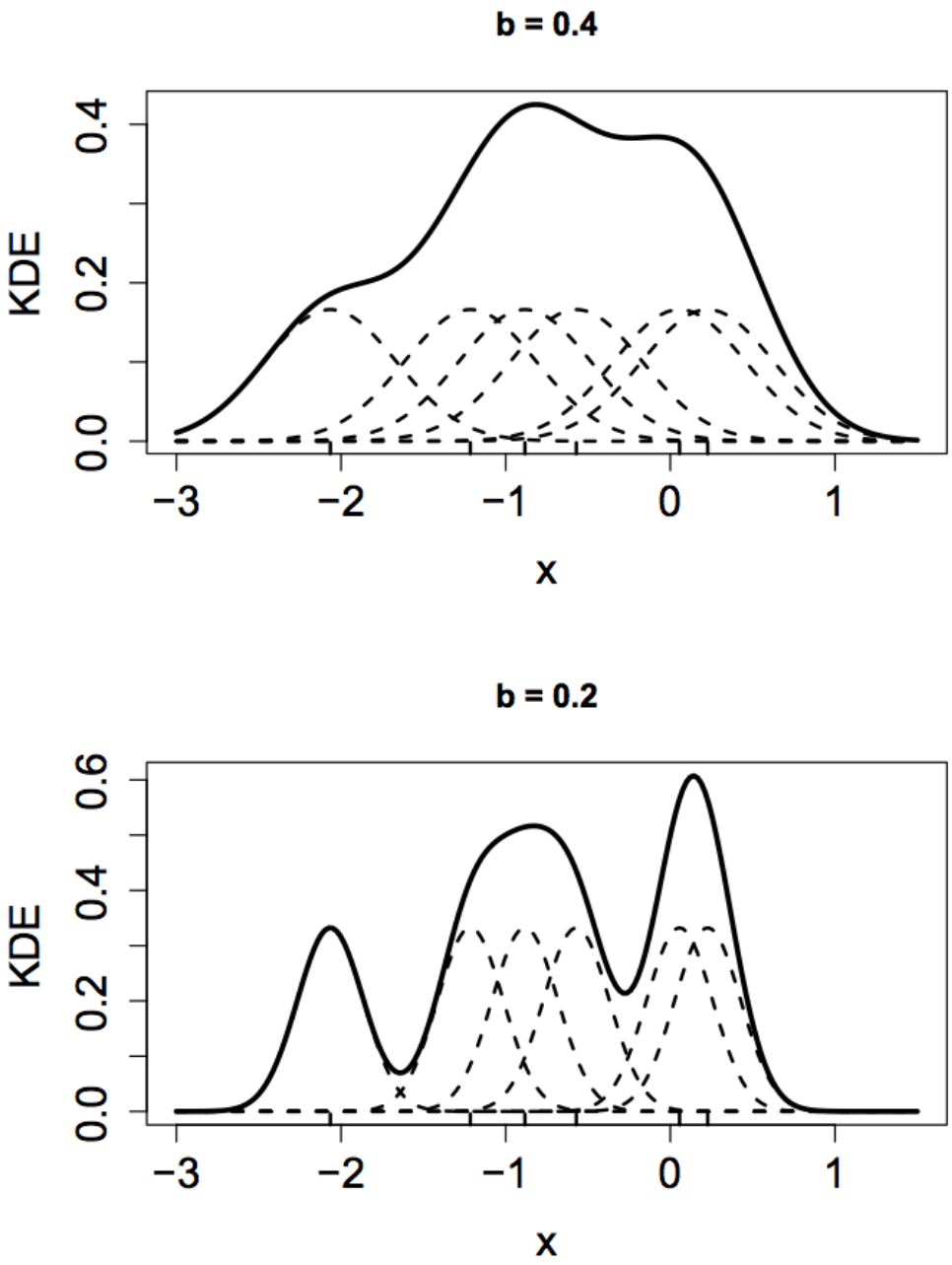
$$\widehat{f(y)} = \frac{1}{nb} \sum_{i=1}^{n} K\left(\frac{Y_i - y}{b}\right).$$

2.1.7 Kernel Density Estimation

The KDE superimposes a density function (the kernel) over each data observation.

- The bandwidth parameter b dictates the width of the kernel.
- Larger values of b mean that the kernels of adjacent observations have a larger effect on the density estimate at a particular observation, y_i.
- In this fashion, b dictates the amount of data *smoothing*.

2.1.8 Illustration of KDE Estimator



This plot was taken directly from Ruppert (2011).

2.1.9 Kernel Density Bandwidth

Choosing b requires a tradeoff between bias and variance.

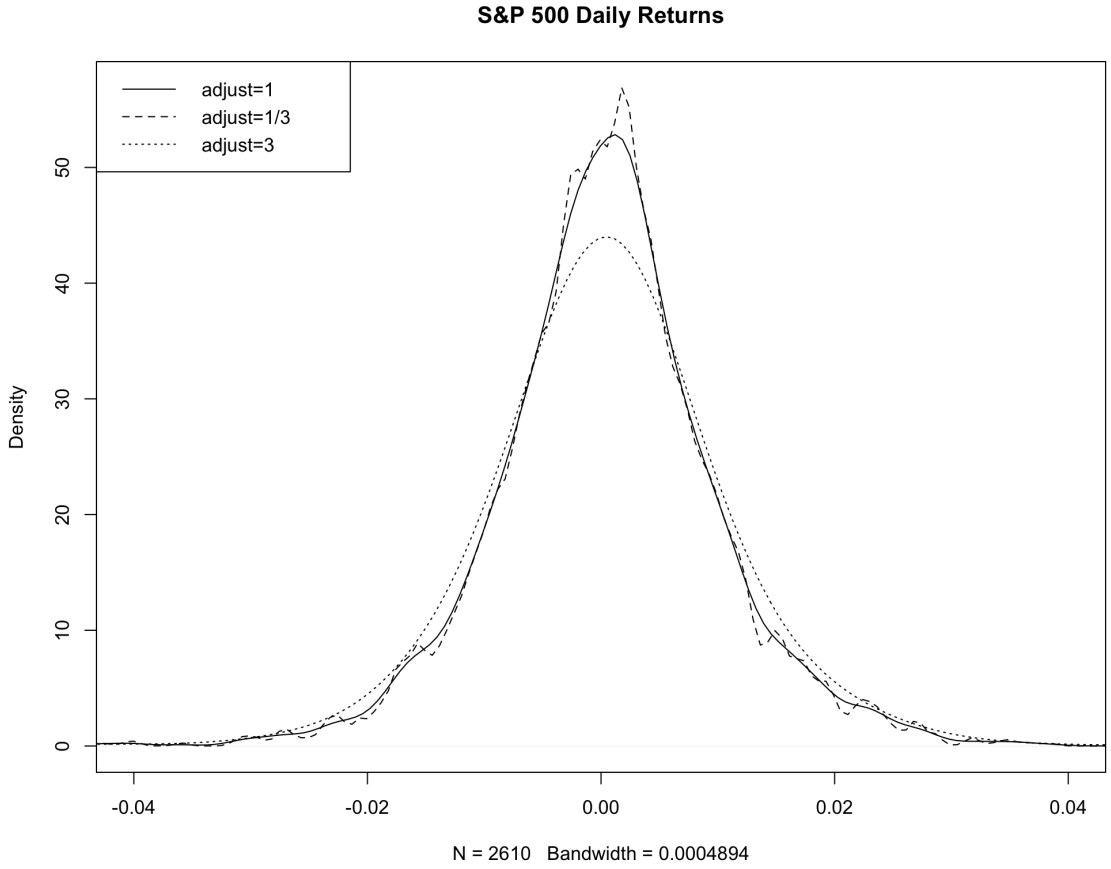
- Small values of b detect fine features of the true density but permit a lot of random variation.
    - The KDE has high variance and low bias.

- If b is too small, the KDE is *undersmoothed* or *overfit* - it adheres too closely to the data.

2.1.10 Kernel Density Bandwidth

- Large values of b smooth over random variation but obscure fine details of the distribution.
  - The KDE has low variance and high bias.
  - If b is too large, the KDE is *oversmoothed* or *underfit* - it misses features of the true density.

2.1.11 KDE of S&P 500 Daily Returns



To create this plot, run the following script:

```
# Kernel density estimates
par(bg="white")
plot(density(spRets, adjust=1/3), lty=2, main="S&P 500 Daily Returns",
             xlim=c(-0.04,0.04))
lines(density(spRets))
lines(density(spRets, adjust=3), lty=3)
legend("topleft", legend=c("adjust=1", "adjust=1/3", "adjust=3"), lty=c(1,2,3))
dev.copy(png, file="sp500KDE.png", height=8, width=10, units='in', res=200)
graphics.off()
```

2.1.12 KDE of S&P 500 Daily Returns

The KDE of the S&P 500 returns suggests a density that resembles a normal distribution.

- We can compare the KDE with a normal distribution with $\mu = \hat{\mu}$ and $\sigma^2 = \hat{\sigma}^2$, where

$$\hat{\mu} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

$$\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{\mu})^2.$$

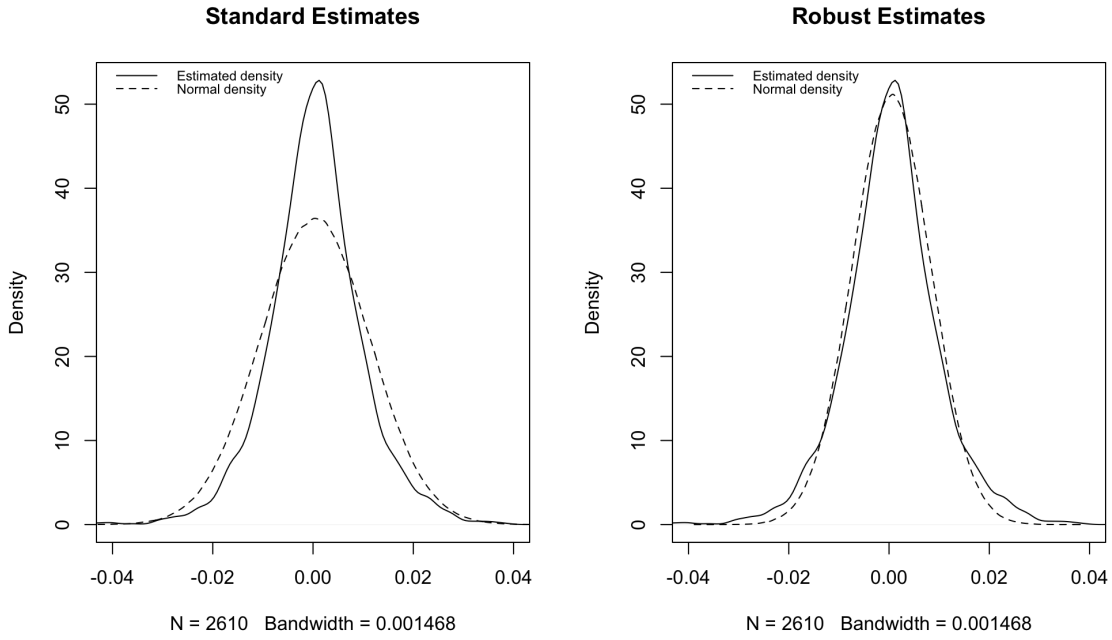
2.1.13 KDE of S&P 500 Daily Returns

- We can also compare the KDE with a normal distribution with $\mu = \tilde{\mu}$ and $\sigma^2 = \tilde{\sigma}^2$

$$\tilde{\mu} = \text{median}\left(\{Y_i\}_{i=1}^{n}\right)$$

$$\tilde{\sigma}^2 = \text{MAD}\left(\{Y_i\}_{i=1}^{n}\right) = \text{median}\left(\{|y_i - \tilde{\mu}|\}_{i=1}^{n}\right).$$

2.1.14 Comparison of KDE with Normal Densities



To create this plot, run the following script:

```
# Comparison of KDE with normals
par(mfrow=c(1,2), bg="white")
normalDens = rnorm(1000000, mean(spRets), sd(spRets))
plot(density(spRets), xlim=c(-0.04,0.04), main="Standard Estimates")
```

```
lines(density(normalDens), lty=2)
legend("topleft", legend=c("Estimated density", "Normal density"),
       lty=c(1,2), cex=0.7, bty='n')
normalDens = rnorm(1000000, median(spRets), mad(spRets))
plot(density(spRets), xlim=c(-0.04,0.04), main="Robust Estimates")
lines(density(normalDens), lty=2)
legend("topleft", legend=c("Estimated density", "Normal density"),
       lty=c(1,2), cex=0.7, bty='n')
dev.copy(png, file="sp500KDENorm.png", height=6, width=10, units='in', res=200)
graphics.off()
```

2.1.15 Comparison of KDE with Normal Densities

Outlying observations in the S&P 500 returns have great influence on the estimates $\hat{\mu}$ and $\hat{\sigma}^2$.

- As a result, a $N(\hat{\mu}, \hat{\sigma})$ deviates substantially from the KDE.

The median, $\tilde{\mu}$, and median absolute deviation, $\tilde{\sigma}^2$, are less sensitive (more robust) to outliers.

- As a result, a $N(\tilde{\mu}, \tilde{\sigma})$ deviates less from the KDE.
- The fit is still not perfect - asset returns are often better approximated with a heavy tailed distribution, like the t.

2.2 Sample Quantiles

2.2.1 Empirical Density Function

Suppose $y_1, y_2, \ldots, y_n$ is a data sample from a true CDF F.

- The *empirical density function* (EDF), $F_n(y)$, is the proportion of the sample that is less than or equal to y:

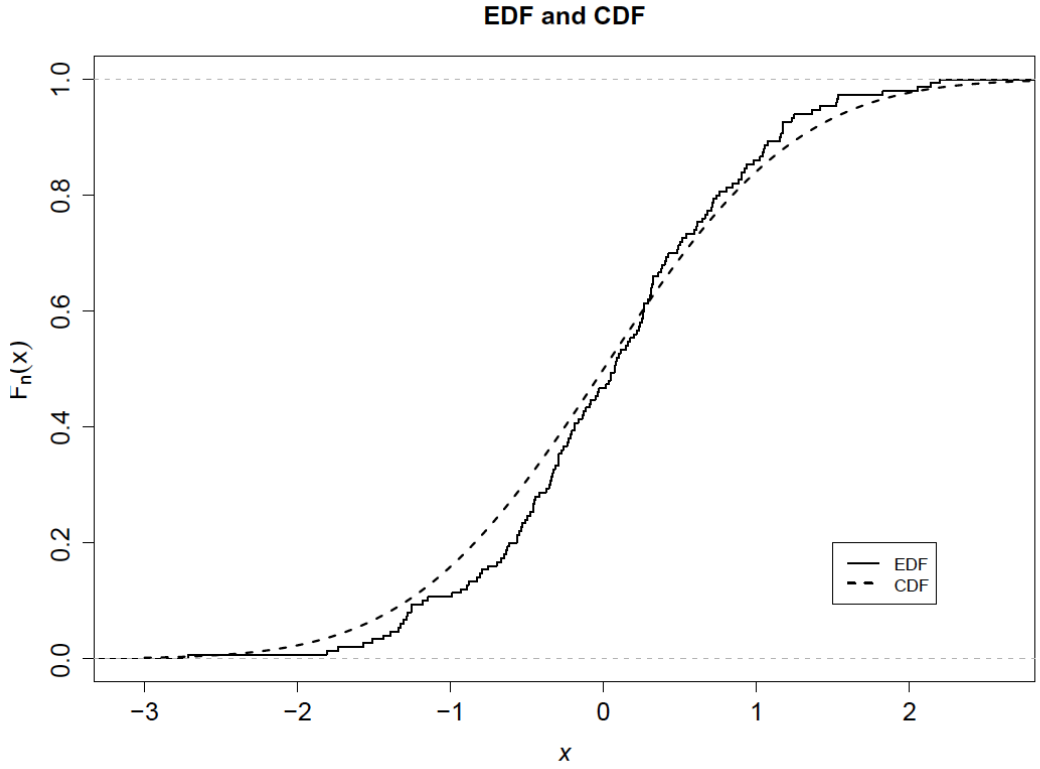
$$F_n(y) = \frac{\sum_{i=1}^{n} I\{y_i \leq y\}}{n},$$

where

$$I\{y_i \leq y\} = \begin{cases} 1 & \text{if } y_i \leq y \\ 0 & \text{otherwise.} \end{cases}$$

2.2.2 EDF Example

The figure below compares the EDF of a sample of 150 observations drawn from a $N(0, 1)$ with the true CDF.



This plot was taken directly from Ruppert (2011).

2.2.3 Order Statistics

Order statistics are the values $y_1, y_2, \ldots, y_n$ ordered from smallest to largest.

- Order statistics are denoted by $y_{(1)}, y_{(2)}, \ldots y_{(n)}$.
- The parentheses in the subscripts differentiate them from the unordered sample.

2.2.4 Quantiles

The q *quantile* of a distribution is the value y such that

$$F_Y(y) = q \quad \Rightarrow \quad y = F_Y^{-1}(q).$$

- Note that $q \in (0, 1)$.
- Quantiles are often called $100q$ th *percentiles*.

2.2.5 Quantiles

Special quantiles:

- Median: $q = 0.5$.
- Quartiles: $q = \{0.25, 0.5, 0.75\}$.

- Quintiles: $q = \{0.2, 0.4, 0.6, 0.8\}$.
- Deciles: $q = \{0.1, 0.2, \ldots, 0.8, 0.9\}$.

2.2.6 Sample Quantiles

The q *sample quantile* of a distribution is the value $y_{(k)}$ such that

$$F_n(y_{(k)}) \leq q.$$

- This is simply the value $y_{(k)}$ where k is qn rounded down to the nearest integer.

2.2.7 Normal Probability Plots

We are often interested in determining whether our data are drawn from a normal distribution.

- If the data is normally distributed, then the q sample quantiles will be approximately equal to $\mu + \sigma\Phi^{-1}(q)$.
  - μ **and** σ **are the true (unobserved) mean and** standard deviation of the data.
  - Φ is the standard normal CDF.

2.2.8 Normal Probability Plots

Hence, a plot of the sample quantiles vs. Φ^{-1} should be roughly linear.

- In practice this is accomplished by plotting $y_{(i)}$ vs. $\Phi(i/(n+1))$, for $i = 1, \ldots, n$.
- Deviations from linearity suggest nonnormality.

2.2.9 Normal Probability Plots

There is no consensus as to which axis should represent the sample quantiles in a normal probability plot.

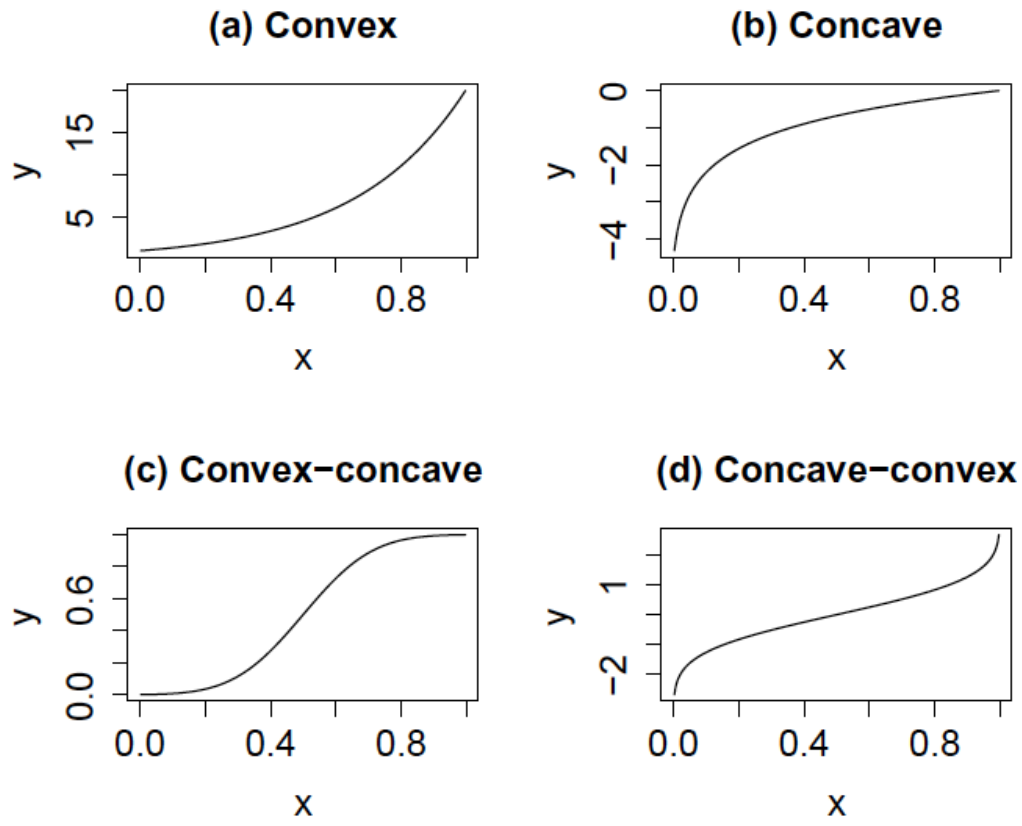
- Interpretation of the plot will depend on the choice of axes for sample and theoretical quantiles.
- We will always place sample quantiles on the x axis.
- In R, the argument 'datax' of the function 'qqnorm' must be set to 'TRUE' (by default, it is 'FALSE').

2.2.10 Interpreting Normal Probability Plots

With the sample quantiles on the x axis:

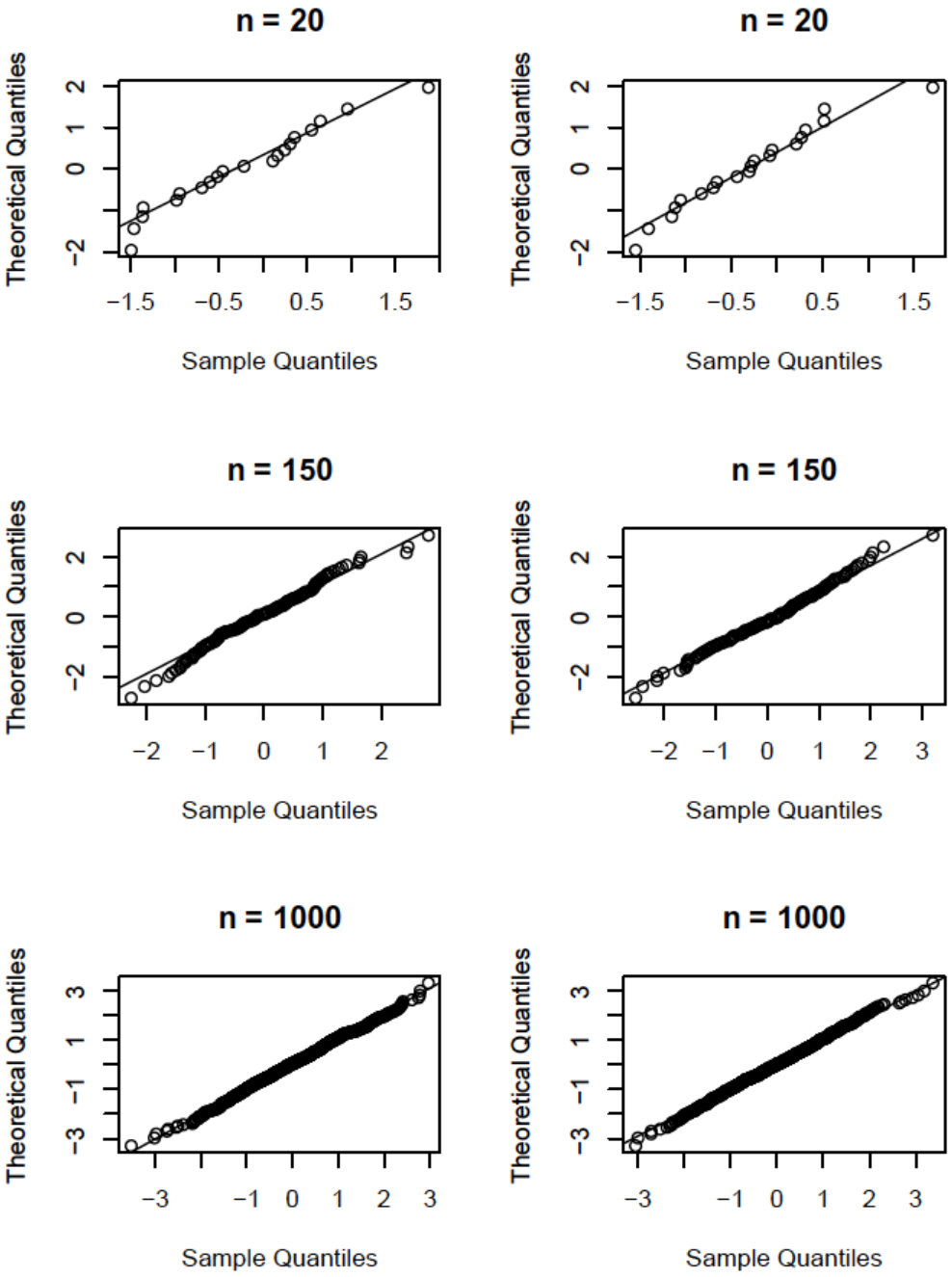
- A convex curve indicates *left skewness*.
- A concave curve indicates *right skewness*.
- A convex-concave curve indicates *heavy tails*.
- A concave-convex curve indicates *light tails*.

2.2.11 Normal Probability Plot Illustrations



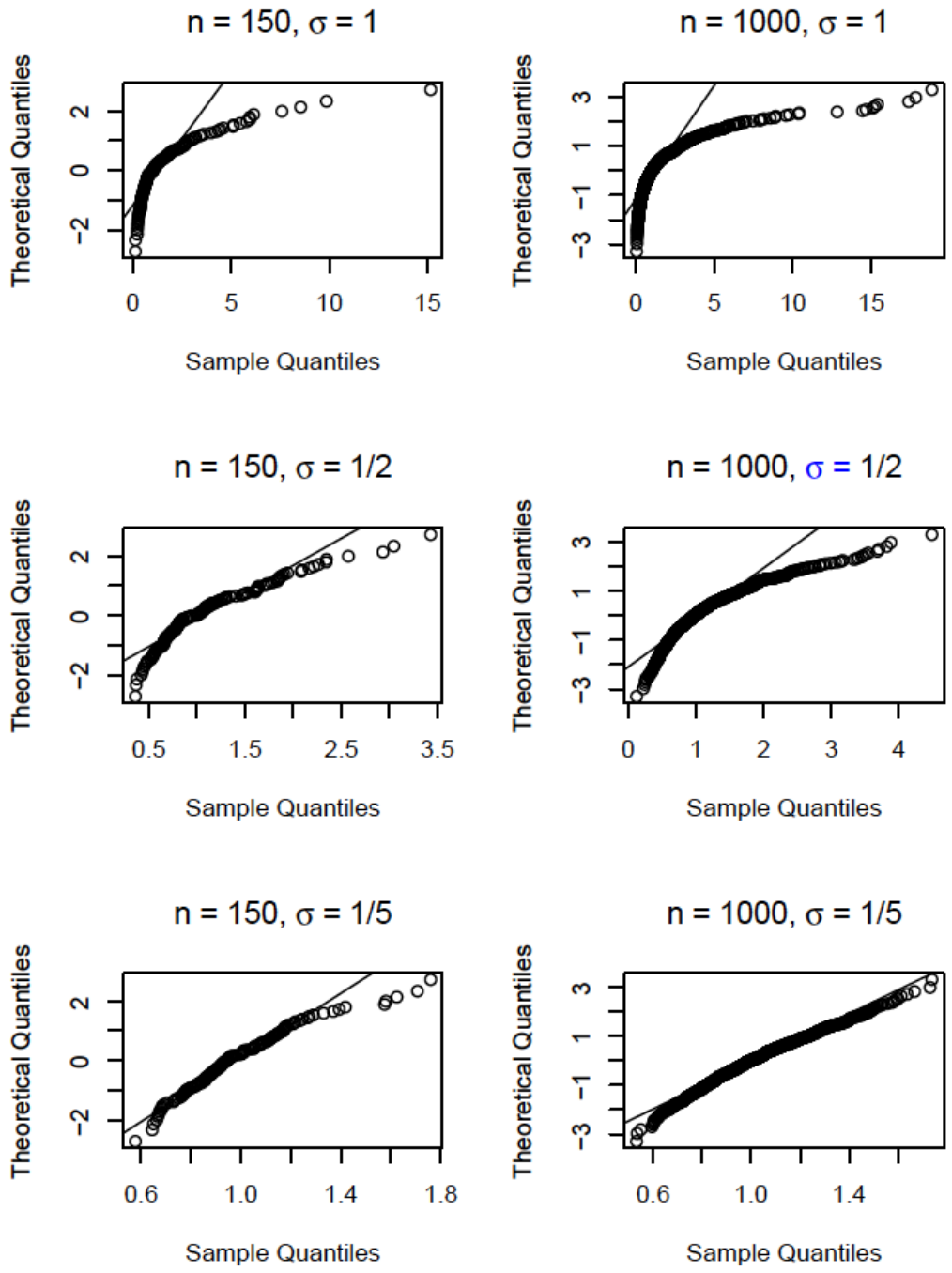
This plot was taken directly from Ruppert (2011).

2.2.12 Normal Prob. Plots for Normal Data



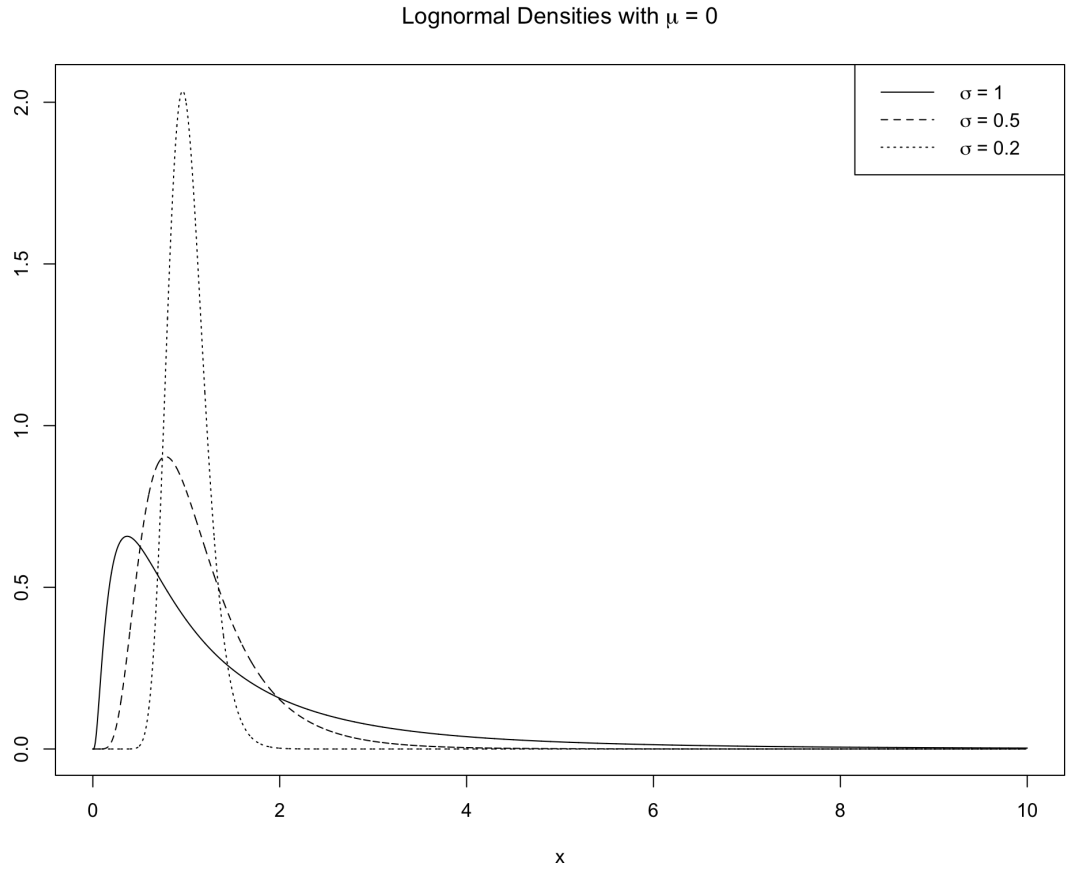
This plot was taken directly from Ruppert (2011).

2.2.13 Normal Prob. Plots for Lognormal Data



This plot was taken directly from Ruppert (2011).

2.2.14 Plots of Lognormal Densities

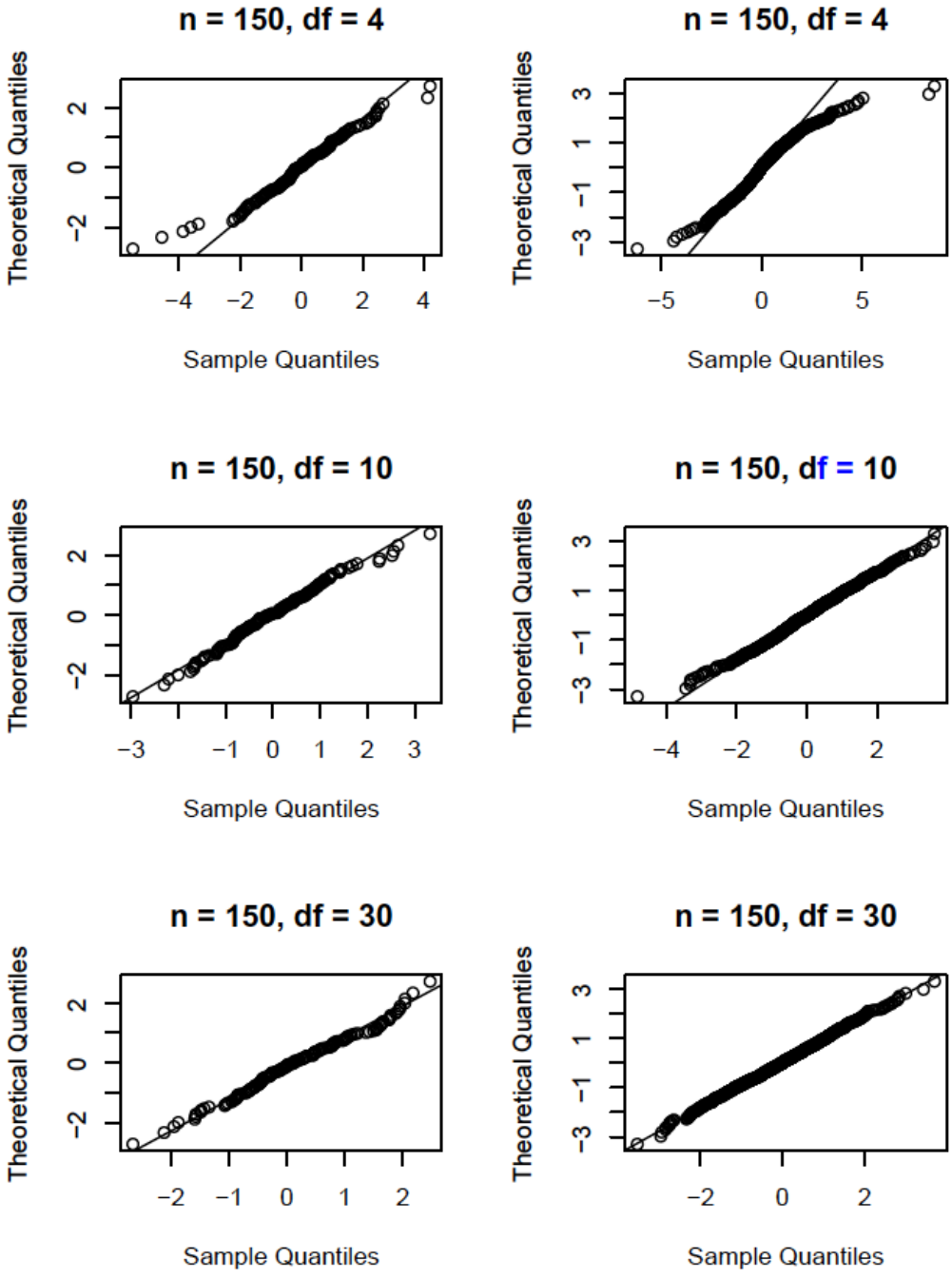


To create this plot, run the following script:

```
##############################################################################
# Plot the Lognormal distribution for mu=0, sigma = 1, 0.5, 0.2
# If X is lognormal(mu, sigma), the ln(X) is normal(mu, sigma)
##############################################################################

# Specify the range of possible random variable values
support = seq(0,10,length=10000);

# Compute the density for each of the lognormals at each x value
d1 = dlnorm(support, 0, 1);
d2 = dlnorm(support, 0, 0.5);
d3 = dlnorm(support, 0, 0.2);

# Plot the lognormals
ymax = max(max(d1), max(d2), max(d3));
par(bg="white")
plot(support, d1, ylim=c(0,ymax), type='l', ylab="", xlab="x",
     main=expression(paste("Lognormal Densities with ", mu, " = 0", sep="")));
lines(support, d2, lty=2);
lines(support, d3, lty=3);
legend("topright", legend=c(expression(paste(sigma, " = 1", sep="")),
       expression(paste(sigma, " = 0.5", sep="")),
```

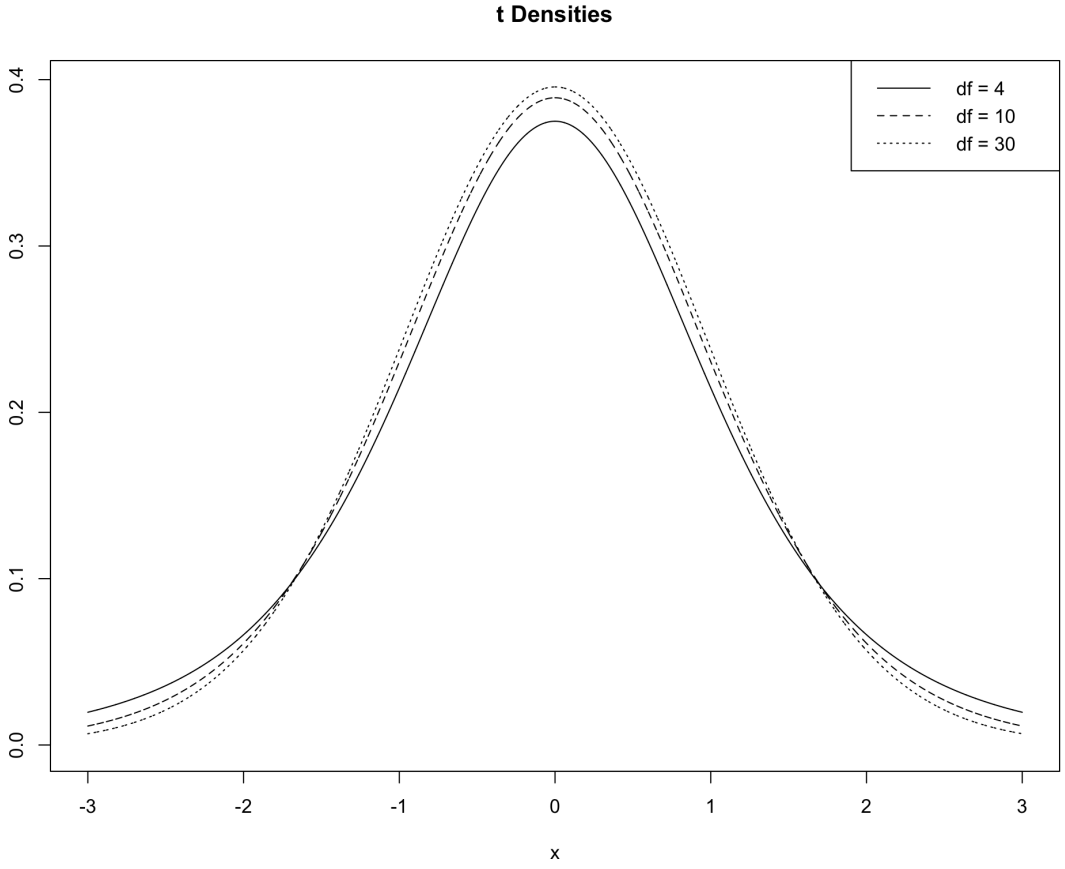
```
      expression(paste(sigma, " = 0.2", sep=""))), lty=c(1,2,3))
dev.copy(png, file="lognormExamples.png", height=8, width=10, units='in', res=200)
graphics.off()
```

2.2.15 Normal Prob. Plots for t Data



This plot was taken directly from Ruppert (2011).

2.2.16 Plots of t Densities

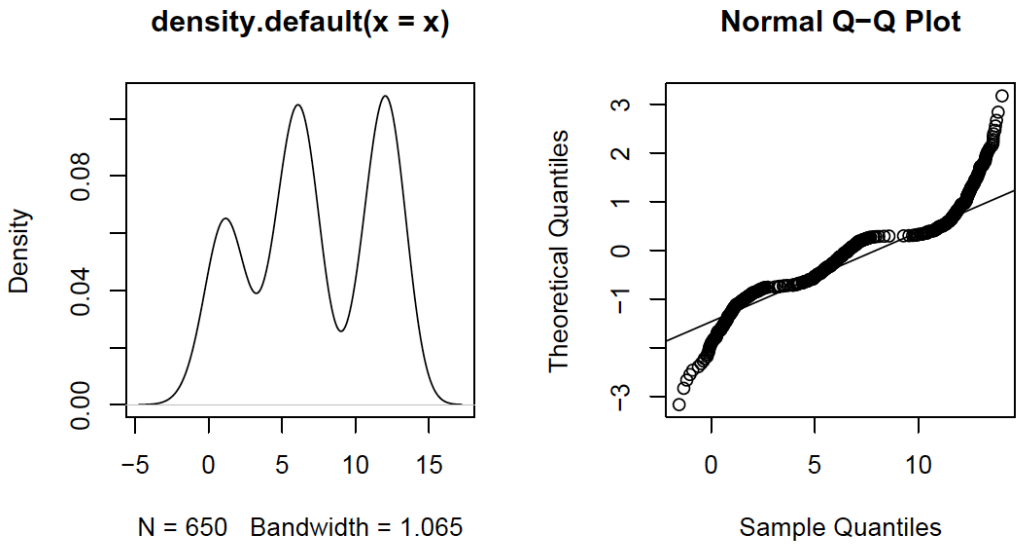


To create this plot, run the following script:

```
################################################################################
# Plot the t distribution for df = 4, 10, 30
################################################################################

# Specify the range of possible random variable values
support = seq(-3,3,length=10000);

# Compute the density for each of the lognormals at each x value
d1 = dt(support, 4);
d2 = dt(support, 10);
d3 = dt(support, 30);

# Plot the t distributions
ymax = max(max(d1), max(d2), max(d3));
par(bg="white")
plot(support, d1, ylim=c(0,ymax), type='l', ylab="", xlab="x", main="t Densities");
lines(support, d2, lty=2);
lines(support, d3, lty=3);
legend("topright", legend=c("df = 4", "df = 10", "df = 30"), lty=c(1,2,3))
dev.copy(png, file="tdistExamples.png", height=8, width=10, units='in', res=200)
graphics.off()
```

2.2.17 Normal Probability Plots vs. KDEs

If the relationship between theoretical and sample quantiles is complex, the KDE is a better tool to understand deviations from normality.



This plot was taken directly from Ruppert (2011).

2.2.18 QQ Plots

Normal probability plots are special examples of *quantile-quantile* (QQ) plots.

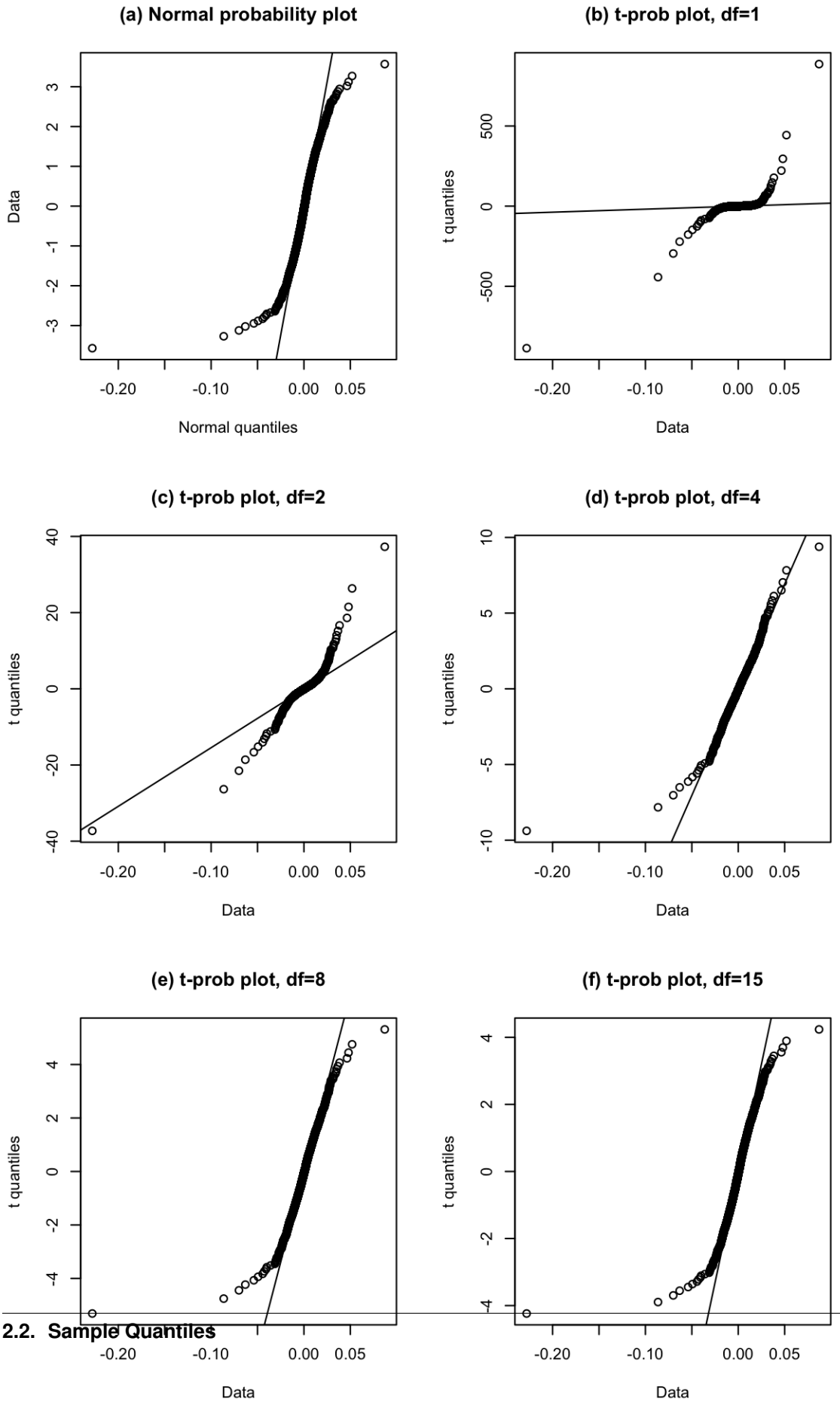
- A QQ plot is a plot of quantiles from one sample or distribution against the quantiles of another sample or distribution.
- For example, we could compare the quantiles of S&P 500 daily returns against a t distribution.
- Alternatively, we could compare the quantiles of S&P 500 daily returns against other financial data.
- QQ plots of multiple datasets indicate which distribution is more/less left/right skewed or which has heavier/lighter tails.

2.2.19 S&P 500 Returns and t Distributions

To create the next two QQ plots, load the data by running the following script.:

```
##############################################################################
# QQ plots
##############################################################################

# Install the Ecdat package if not already installed
#install.packages("Ecdat")
library(Ecdat);

# Get the DM/Dollar data
data(Garch);
dm = Garch$dm;
```

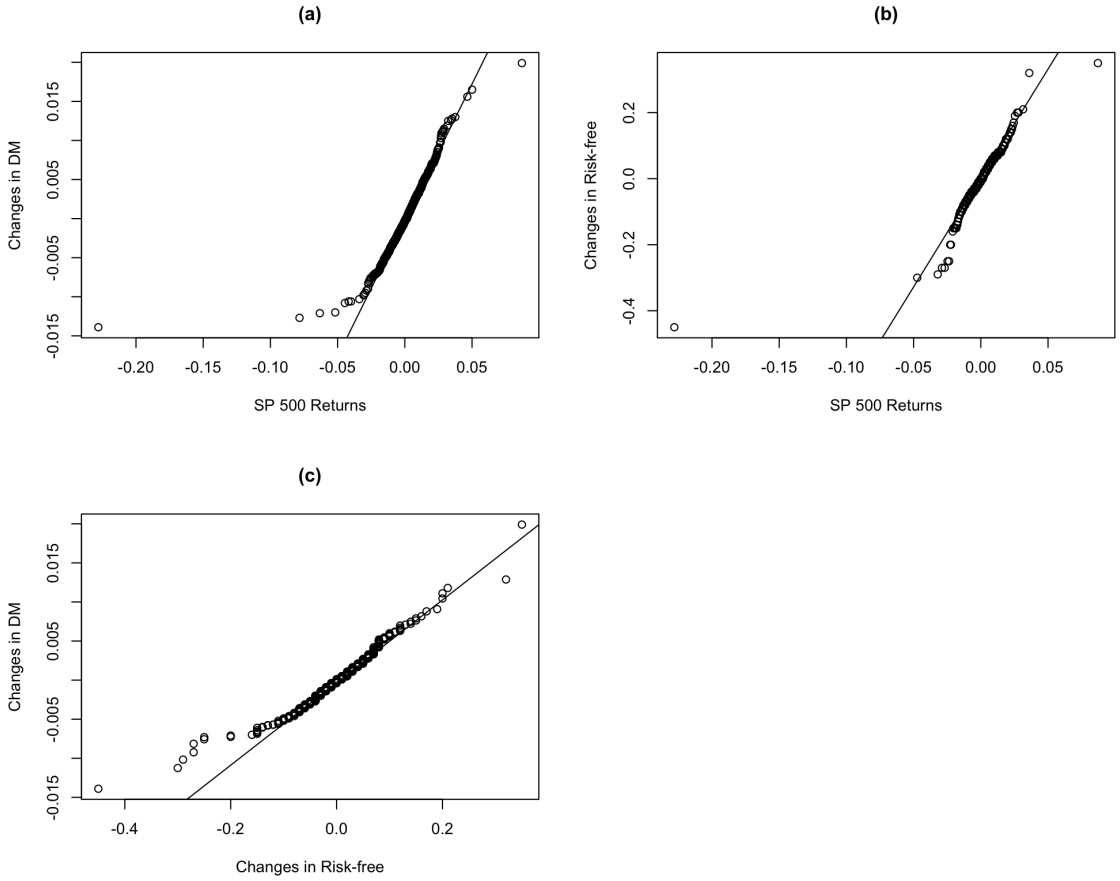
```
# Get risk-free rate data
data(Capm);
rf = Capm$rf;

# Get SP500 data
data(SP500);
r500 = SP500$r500
```


(a) Normal probability plot
Data
Normal quantiles
(b) t-prob plot, df=1
t quantiles
Data
(c) t-prob plot, df=2
t quantiles
Data
(d) t-prob plot, df=4
t quantiles
Data
(e) t-prob plot, df=8
t quantiles
Data
(f) t-prob plot, df=15
t quantiles
Data

To create this plot, run the following script:

```
# QQ plots of SP500 returns against t distributions
n = length(r500);
qGrid = (1:n)/(n+1);
par(mfrow=c(3,2), bg="white")
qqnorm(r500, datax=TRUE, xlab="Data", ylab="Normal quantiles",
       main="(a) Normal probability plot")
qqline(r500, datax=TRUE)
qqplot(r500, qt(qGrid, df=1), xlab="Data", ylab="t quantiles",
       main="(b) t-prob plot, df=1")
qqline(r500, datax=TRUE, distribution=function(p){qt(p, df=1)})
qqplot(r500, qt(qGrid, df=2), xlab="Data", ylab="t quantiles",
       main="(c) t-prob plot, df=2")
qqline(r500, datax=TRUE, distribution=function(p){qt(p, df=2)})
qqplot(r500, qt(qGrid, df=4), xlab="Data", ylab="t quantiles",
       main="(d) t-prob plot, df=4")
qqline(r500, datax=TRUE, distribution=function(p){qt(p, df=4)})
qqplot(r500, qt(qGrid, df=8), xlab="Data", ylab="t quantiles",
       main="(e) t-prob plot, df=8")
qqline(r500, datax=TRUE, distribution=function(p){qt(p, df=8)})
qqplot(r500, qt(qGrid, df=15), xlab="Data", ylab="t quantiles",
       main="(f) t-prob plot, df=15")
qqline(r500, datax=TRUE, distribution=function(p){qt(p, df=15)})
dev.copy(png, file="sp500QQ.png", height=10, width=6, units='in', res=200)
graphics.off()
```

2.2.20 S&P 500 Returns, DM/Dollar and Risk-free



To create this plot, run the following script:

```
# QQ plots of SP500, DM and risk-free
par(mfrow=c(2,2), bg="white")
qqplot(r500, diff(dm), xlab="SP 500 Returns", ylab="Changes in DM", main="(a)")
abline(lm(quantile(diff(dm), c(0.25,0.75))~quantile(r500, c(0.25,0.75))))
qqplot(r500, diff(rf), xlab="SP 500 Returns", ylab="Changes in Risk-free", main="(b)")
abline(lm(quantile(diff(rf), c(0.25,0.75))~quantile(r500, c(0.25,0.75))))
qqplot(diff(rf), diff(dm), xlab="Changes in Risk-free", ylab="Changes in DM",
       main="(c)")
abline(lm(quantile(diff(dm), c(0.25,0.75))~quantile(diff(rf), c(0.25,0.75))))
dev.copy(png, file="sp-dm-rf-qq.png", height=8, width=10, units='in', res=200)
graphics.off()
```

2.3 Data Transformations

2.3.1 Transformations

Data often deviate from normality and exhibit characteristics (skewness, kurtosis) that are difficult to model.

- Transforming data using some functional form will often result in observations that are easier to model.

- The most typical transformations are the natural logarithm and the square root.

2.3.2 Logarithmic Transformation

Given independent and dependent variables, (x_t, y_t), the natural logarithm transformation is appropriate under several circumstances:

- y_t is strictly positive (the log of a negative number does not exist).
- y_t increases exponentially (faster than linearly) as x_t increases.
- The variance in y_t appears to depend on its mean (heteroskedasticity).

2.3.3 Logarithmic Transformation

Consider the relationship

$$y_t = \exp(\beta x_t)\exp(\epsilon_t),$$

where $\epsilon_t \sim \mathcal{N}(0, \sigma)$.

- If $\epsilon_t \sim \mathcal{N}(0, \sigma)$, then $\exp(\epsilon_t) \sim \mathcal{LN}(0, \sigma)$.
- In this case,

$$E\left[\exp(\epsilon_t)\right] = \exp(0.5\sigma^2)$$

$$Var\left(\exp(\epsilon_t)\right) = \left(\exp(\sigma^2) - 1\right)\exp(\sigma^2).$$

2.3.4 Logarithmic Transformation

Thus,

$$E[y_t] = \exp(\beta x_t)\exp(0.5\sigma^2)$$

$$Var(y_t) = \exp(2\beta x_t)\left(\exp(\sigma^2) - 1\right)\exp(\sigma^2).$$

- That is, $E[y_t]$ grows exponentially with x_t and $Var(y_t)$ is heteroskedastic.

2.3.5 Logarithmic Transformation

Taking the natural logarithm

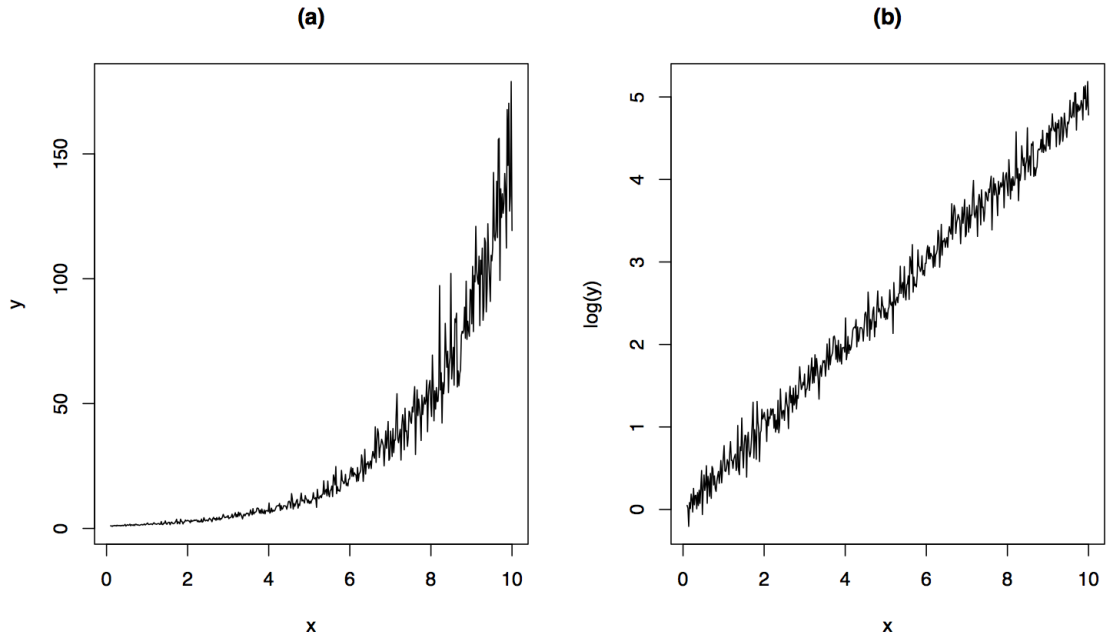
$$\log(y_t) = \beta x_t + \epsilon_t,$$

- $E\left[\log(y_t)\right]$ grows linearly with x_t.
- $Var\left(\log(y_t)\right)$ is homoskedastic.

2.3.6 Logarithmic Transformation Example

Given, $\beta = 0.5$ and $\epsilon_t \sim \mathcal{N}(0, 0.15)$, the plot below depicts

$$y_t = \exp(\beta x_t)\exp(\epsilon_t)$$
$$\log(y_t) = \beta x_t + \epsilon_t.$$

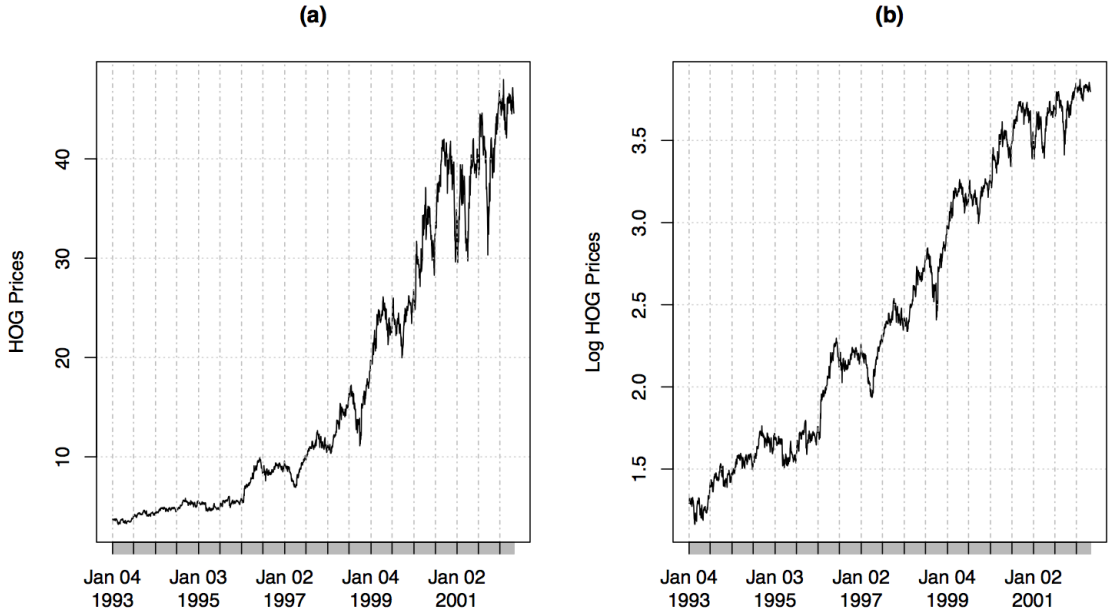


To create this plot, run the following script:

```
setwd("~/Dropbox/Academics/Teaching/Econ114/S2013/RScripts/")
library(quantmod)

# Number of observations
n = 500;

# Parameters driving the true relationship
beta = 0.5;
sigma = 0.15

# The independent variable
x = seq(0.1,10,length=n)

# Random shocks drawn for a normal
eps = rnorm(n,0,sigma);

# Create the independent variable based on specified model
y = exp(beta*x)*exp(eps)

# Plot levels of y and log of y
pdf(file="logTransExample.pdf", height=6, width=10)
par(mfrow=c(1,2))
plot(x, y, type='l', main="(a)")
plot(x, log(y), type='l', main="(b)")
graphics.off()
```

2.3.7 Logarithmic Transformation Example

Asset prices often display the characteristics that are suitable for a logarithmic transformation.



To create this plot, run the following script:

```
getSymbols("HOG", from="1993-01-01", to="2002-04-30")
pdf(file="hogTransExample.pdf", height=6, width=10)
par(mfrow=c(1,2))
plot(HOG$HOG.Adj, ylab="HOG Prices", main="(a)")
plot(log(HOG$HOG.Adj), ylab="Log HOG Prices", main="(b)")
graphics.off()
```

2.3.8 Box-Cox Power Transformations

Generally speaking, the set of transformations

$$y^\alpha = \begin{cases} \frac{y^\alpha - 1}{\alpha} & \alpha \neq 0 \\ \log(y) & \alpha = 0, \end{cases}$$

Is known as the family of *Box-Cox power transformations*.

2.3.9 Correcting Skewness and Heteroskedasticity

Suppose a set of data observations, y_t, appear to be right skewed and have variance increasing with it's mean.

- A concave transformation with $\alpha < 1$ will reduce the skewness and stabilize the variance.
- The smaller the value of α, the greater the effect of the transformation.
- Selecting $\alpha < 1$ too small may result in left skewness or variance decreasing with the mean (or both).
- The α that creates the most symmetric data may not be the best for stabilizing variance - there may be a tradeoff.

2.3.10 Box Cox Example

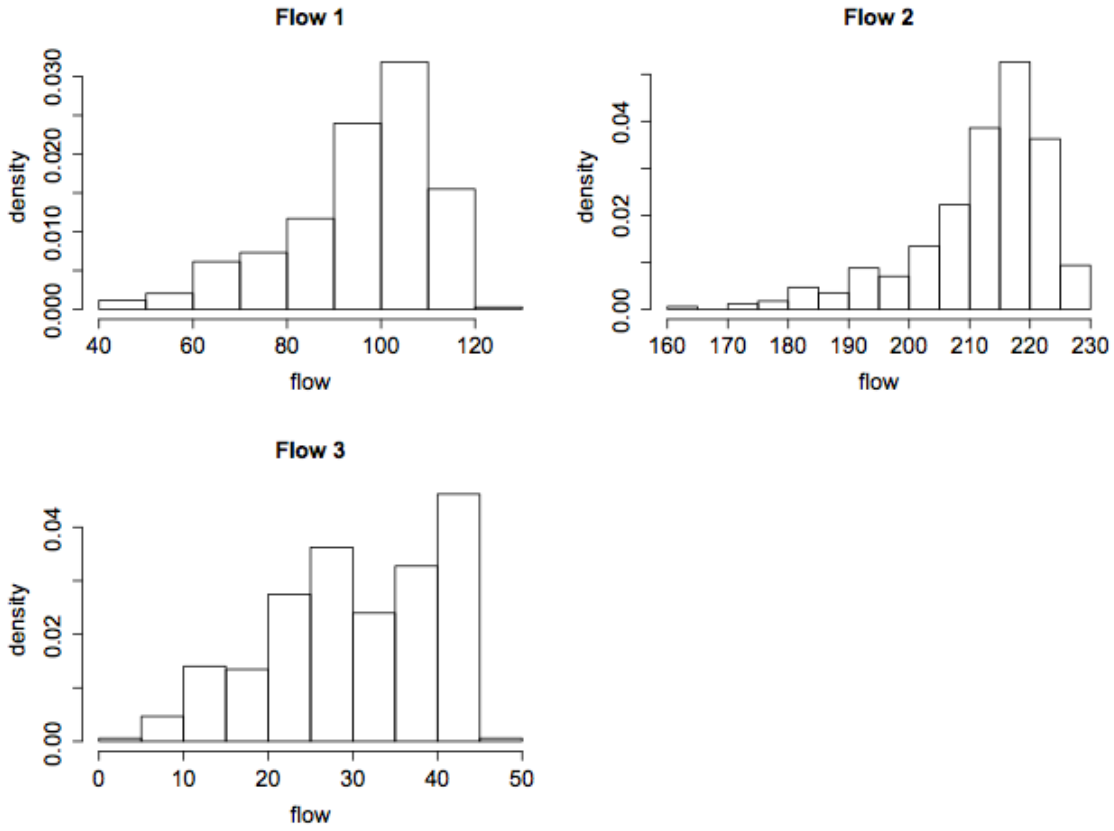


Fig. 4.20. *Histograms of daily flows in three pipelines.*

This plot was taken directly from Ruppert (2011).

2.3.11 Box Cox Example

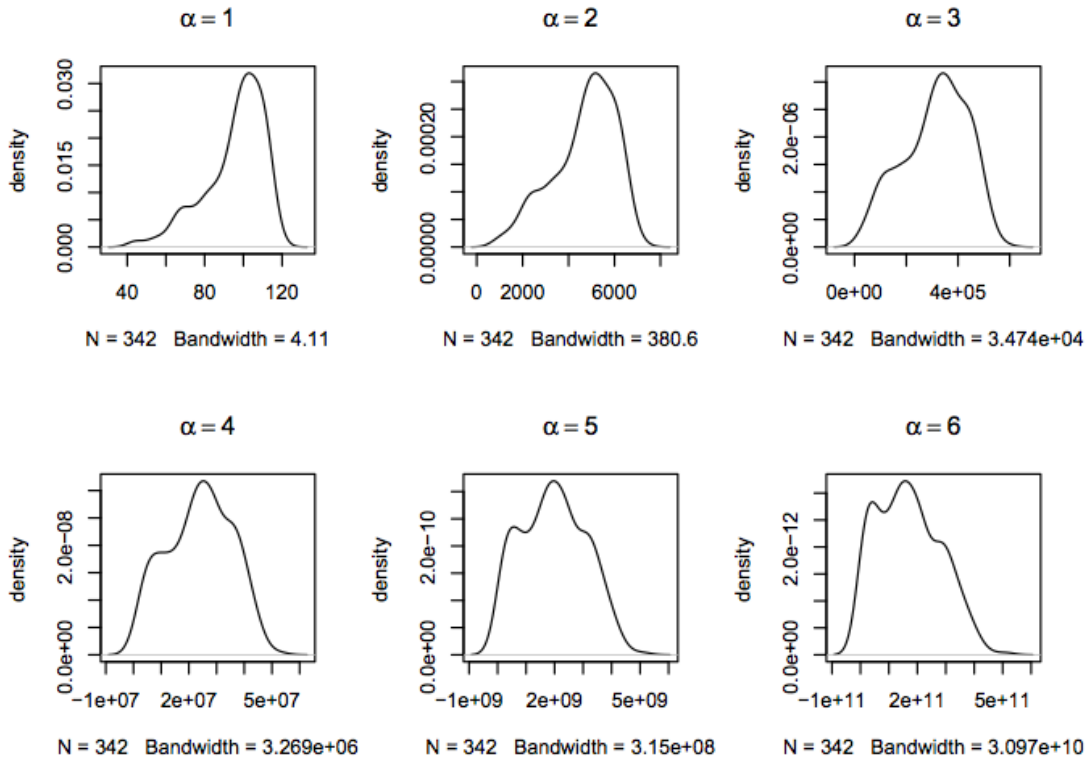


Fig. 4.21. *Kernel density estimates for gas flows in pipeline 1 with Box–Cox transformations.*

This plot was taken directly from Ruppert (2011).

2.3.12 Geometry of Transformations

Transformations can be beneficial because they stretch observations apart in some regions and push them together in other regions.

- If data are right skewed, then a concave transformation will
  - Stretch the distances between observations at the lower end of the distribution.
  - Compress the distances between observations at the upper end of the distribution.
- The degree of stretching and compressing will depend on the derivatives of the transformation function.

2.3.13 Geometry of Transformations

For two values, x and x' close to each other, Taylor's theorem says

$$|h(x) - h(x')| \approx h'(x)|x - x'|.$$

- $h(x)$ and $h(x')$ will be pushed apart where $h'(x)$ is large.
- $h(x)$ and $h(x')$ will be pushed together where $h'(x)$ is small.

- $h'(x)$ is a decreasing function of x if h is concave.

2.3.14 Geometry of Transformations

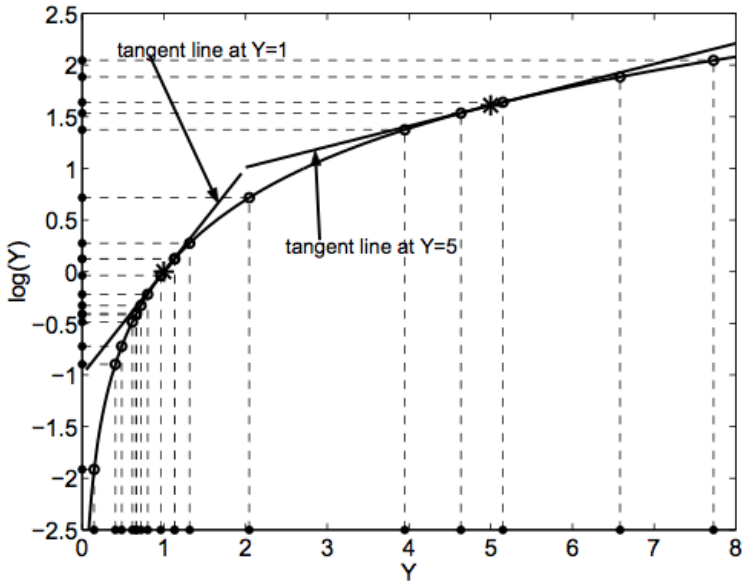


Fig. 4.23. *A symmetrizing transformation. The skewed lognormal data on the horizontal axis are transformed to symmetry by the log transformation.*

This plot was taken directly from Ruppert (2011).

2.3.15 Geometry of Transformations

Similarly, if the variance of a data set increases with its mean, a concave transformation will

- Push more variable values closer together (for large values of the data).
- Push less variable values further apart (for small values of the data).

2.3.16 Geometry of Transformations

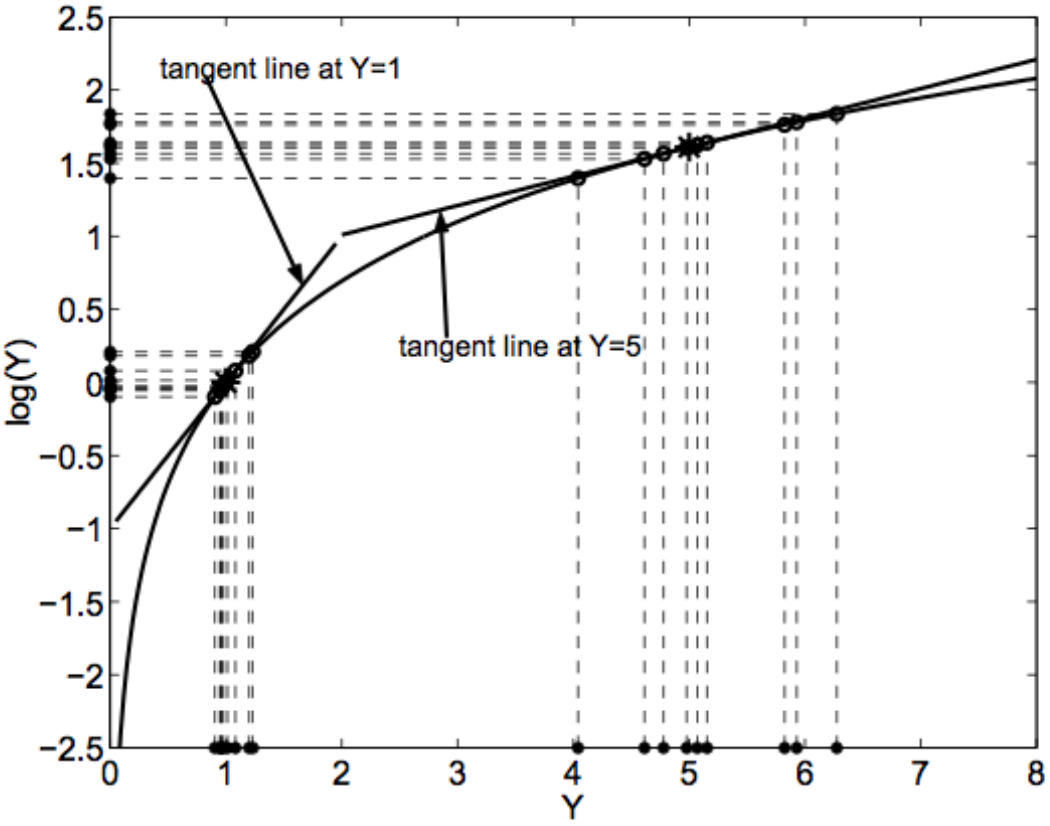


Fig. 4.24. *A variance-stabilizing transformation.*

This plot was taken directly from Ruppert (2011).

CHAPTER

THREE

DISTRIBUTIONS

Contents:

3.1 Moments

3.1.1 Moments

Given a random variable X:

- The k -th *moment* of X is

$$E[X^k] = \int_{-\infty}^{\infty} x^k f(x) dx.$$

- The k -th *central moment* of X is

$$\mu_k = E\left[(X - E[X])^k\right] = \int_{-\infty}^{\infty} (x - E[X])^k f(x) dx.$$

3.1.2 Moments

Some special cases for *any* random variable X:

- The first moment of X is its mean.
- The first *central* moment of X is zero.
- The second *central* moment of X is its variance.

3.1.3 Sample Moments

Given realizations $x_1, \ldots, x_n$ of a random variable X,

- The moments of X can be approximated by replacing expectations with simple averages:

$$\hat{\mu}_k = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^k, \quad \text{where} \quad \bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i.$$

- For example, the sample variance is

$$\hat{\mu}_2 = \hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2.$$

3.1.4 Skewness

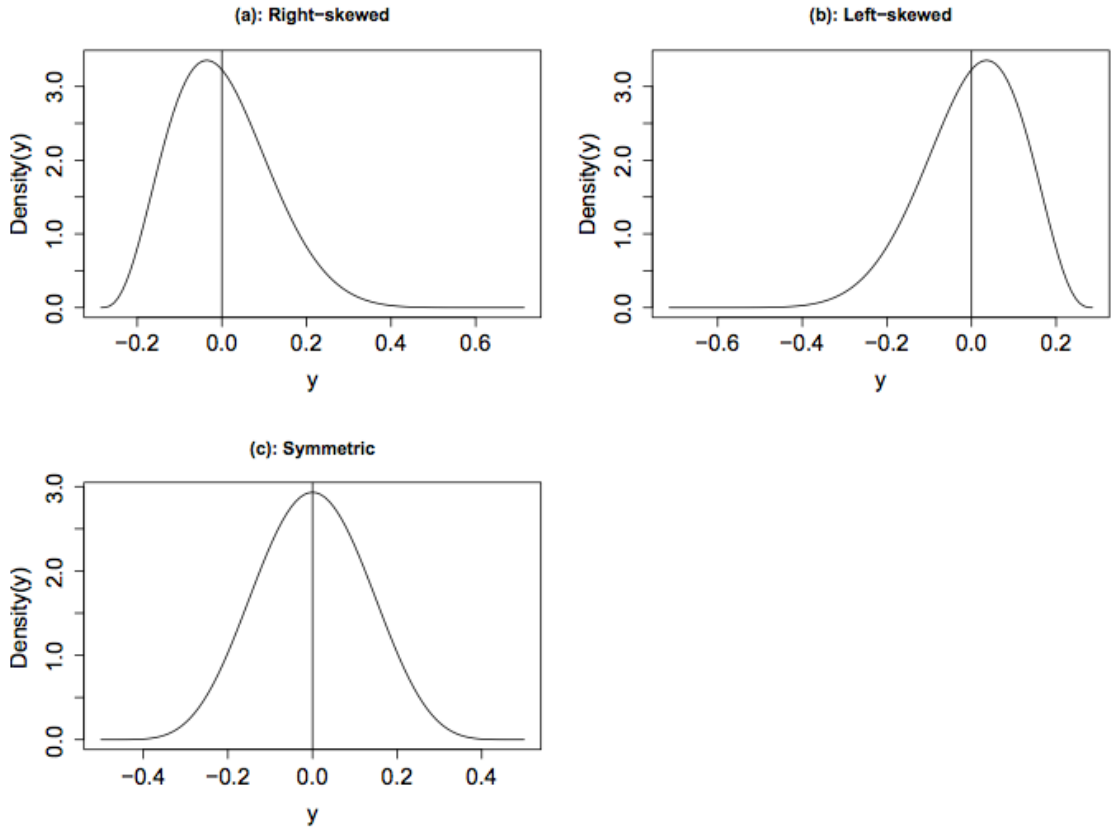
Skewness measures the degree of asymmetry of a distribution.

- Formally, skewness is defined as

$$Sk = E\left[\left(\frac{X - E[X]}{\sigma}\right)^3\right] = \frac{\mu_3}{\sigma^3}.$$

- Zero skewness corresponds to a symmetric distribution.
- Positive skewness indicates a relatively long right tail.
- Negative skewness indicates a relatively long left tail.

3.1.5 Skewness Example



This plot was taken directly from Ruppert (2011).

3.1.6 Kurtosis

Kurtosis measures the extent to which probability is concentrated in the center and tails of a distribution rather than the "shoulders".

- Formally, kurtosis is defined as

$$Kur = E\left[\left(\frac{X - E[Y]}{\sigma}\right)^4\right] = \frac{\mu_4}{\sigma^4}.$$

- High values of kurtosis indicate heavy tails and low values indicate light tails.

3.1.7 Kurtosis

- For skewed distributions, kurtosis may measure both asymmetry and tail weight.
- For symmetric distributions, kurtosis only measures tail weight.

3.1.8 Kurtosis

The Kurtosis of a *all* normal distributions is 3.

- *Excess Kurtosis*, $Kur - 3$, is a measure of the kurtosis of a distribution relative to that of a normal.
- If excess kurtosis is positive, the distribution has heavier tails than a normal.
- If excess kurtosis is negative, the distribution has lighter tails than a normal.

3.1.9 Kurtosis of t-Distribution

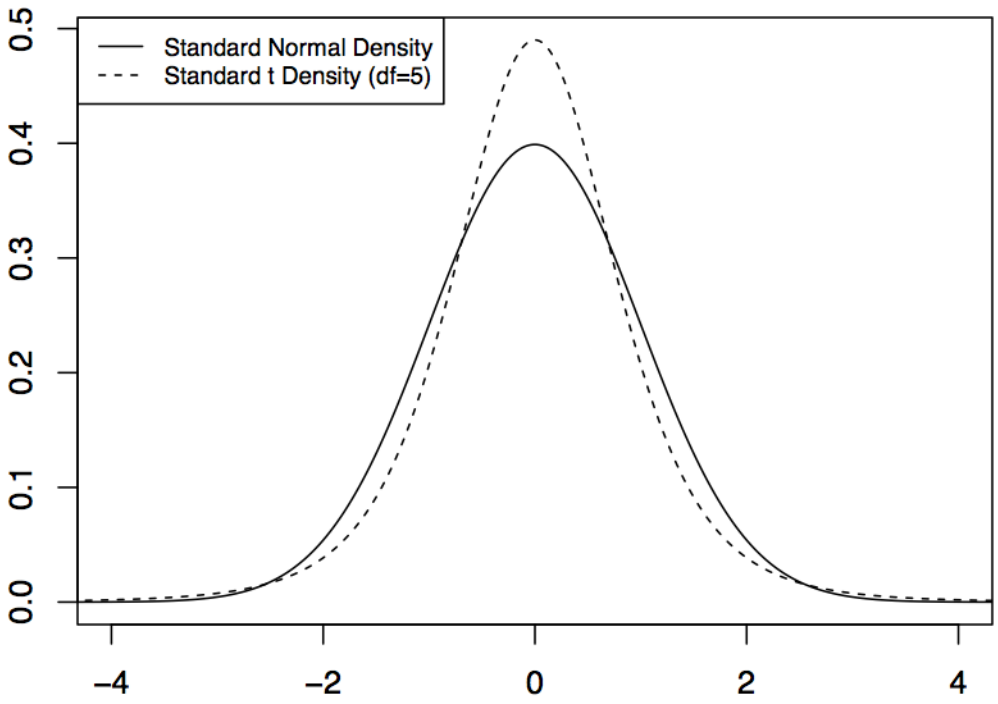
Let t_ν denote a random variable that has a t-distribution with ν degrees of freedom.

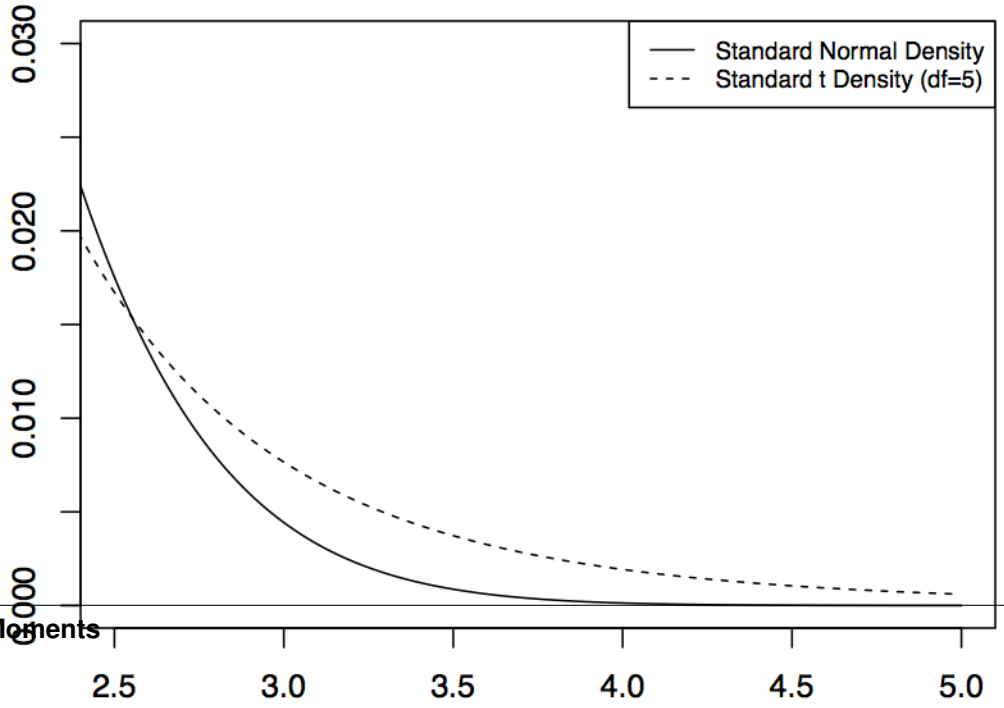
- The kurtosis of t_ν exists only for $\nu > 4$ and is equal to

$$Kur(t_\nu) = 3 + \frac{6}{\nu - 4}.$$

- So, for a t_5-distribution, the kurtosis is 9.
- Clearly, as $\nu \to \infty$, $Kur(t_\nu) \to 3$, which is the kurtosis of a normal.
- This makes sense because $t_\nu \to \mathcal{N}(0, 1)$ as $\nu \to \infty$.

3.1.10 Kurtosis Example





To create this plot, run the following script:

```
setwd("~/Dropbox/Academics/Teaching/Econ114/W2013/RScripts/")

################################################################################
# Comparison of tail weight of Normal and t distributions
################################################################################

# The possible values of the random variables over which to plot densities
xGrid = seq(-5, 5, length=1000)

# Compute values of the normal density at each x
normDens = dnorm(xGrid, 0, 1)

# The degrees of freedom for the t
nu = 5;

# Compute the standardized t density at each x
# Note that to compute the standard t, we must multiply by sqrt(nu/(nu-2))
mult = sqrt(nu/(nu-2))
tDens = mult*dt(mult*xGrid, df=nu)

# Plot the two densities
pdf(file="tNormComp.pdf", height=10, width=6)
par(mfrow=c(2,1))
ymin = min(min(normDens), min(tDens))
ymax = max(max(normDens), max(tDens))
plot(xGrid, normDens, type='l', xlim=c(-4, 4), ylim=c(ymin, ymax), xlab="", ylab="")
lines(xGrid, tDens, lty=2)
legend('topleft', legend=c("Standard Normal Density","Standard t Density (df=5)"),
       lty=c(1,2), cex=0.75)

       # Zoom in on the upper tails
       plot(xGrid, normDens, type='l', xlim=c(2.5, 5), ylim=c(0, 0.03), xlab="", ylab="")
       lines(xGrid, tDens, lty=2)
       legend('topright', legend=c("Standard Normal Density","Standard t Density (df=5)"),
lty=c(1,2), cex=0.75)

graphics.off()
```

3.1.11 Sample Skewness and Kurtosis

Given realizations $x_1, \ldots, x_n$ of a random variable X, skewness and kurtosis can be approximated by

$$\widehat{SK} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\hat{\sigma}}\right)^3$$

$$\widehat{Kur} = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{x_i - \bar{x}}{\hat{\sigma}}\right)^4.$$

3.1.12 Sample Skewness and Kurtosis

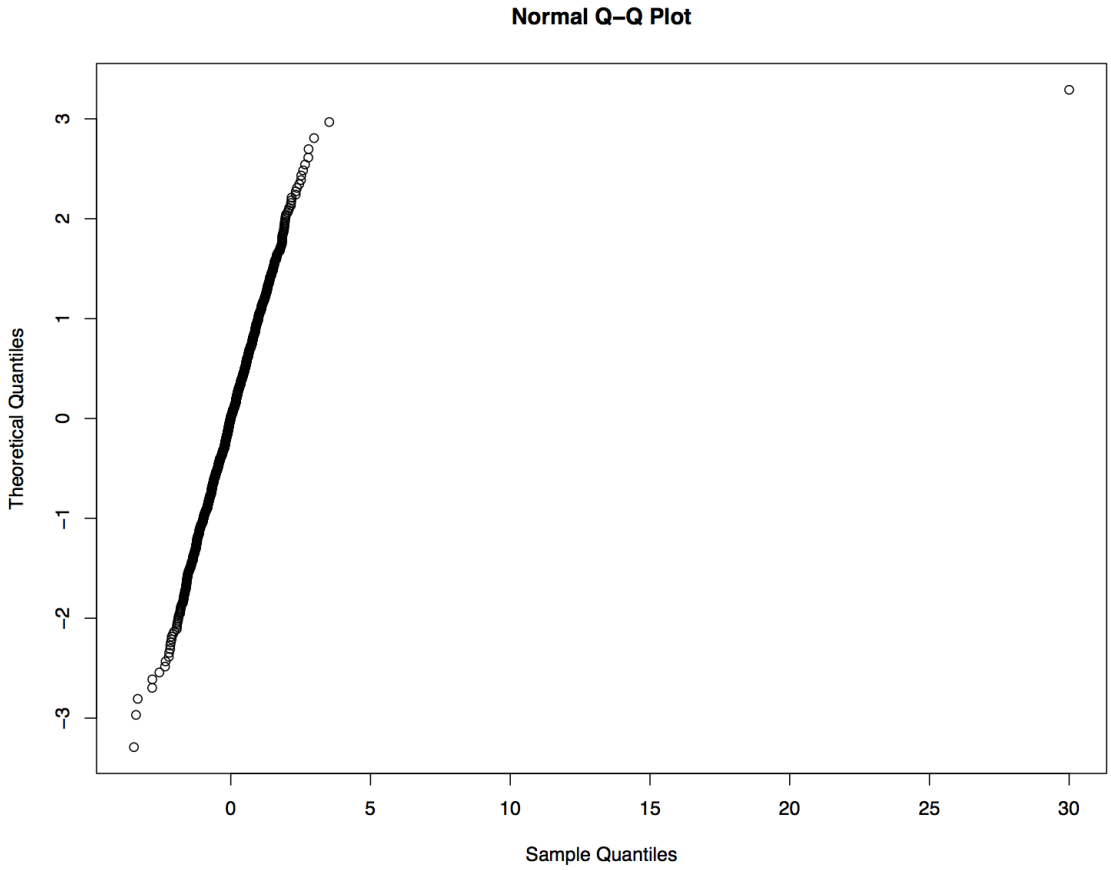
Sample skewness and kurtosis can be used to diagnose normality.

- However, sample skewness and kurtosis are heavily influenced by outliers.

3.1.13 Sample Skewness and Kurtosis

- Consider a random sample of 999 values drawn from a $\mathcal{N}(0, 1)$ distribution.
- The sample skewness and kurtosis are 0.0072 and 3.17, respectively.
- These are close to the true values of 0 and 3.
- If one outlier equal to 30 is added to the dataset, the sample skewness and kurtosis become 10.48 and 231.05, respectively.

3.1.14 Sample Skewness and Kurtosis



To create this plot, run the following script:

```
################################################################################
# Sample Skewness and Kurtosis with Outlier
################################################################################

# Draw 999 values from a N(0,1)
normSamp = rnorm(999, 0, 1)

# Create an outlier
outlier = 30

# Put the normal sample and outlier together in one sample
```

```
totalSamp = c(normSamp, outlier)

# Normal QQ-plot
pdf(file="outlierNormQQ.pdf", height=8, width=10)
qqnorm(totalSamp, datax=TRUE)
graphics.off()

# Compute sample skewness of only normal part
skPart = mean(((normSamp - mean(normSamp))/sd(normSamp))^3)

# Compute sample kurtosis of only normal part
kurPart = mean(((normSamp - mean(normSamp))/sd(normSamp))^4)

# Compute sample skewness
sk = mean(((totalSamp - mean(totalSamp))/sd(totalSamp))^3)

# Compute sample kurtosis
kur = mean(((totalSamp - mean(totalSamp))/sd(totalSamp))^4)
```

3.1.15 Location, Scale and Shape Parameters

- A *location parameter* shifts a distribution to the right or left without changing the distribution's variability or shape.
- A *scale parameter* changes the variability of a distribution without changing its location or shape.
- A parameter is a scale parameter if it increases by $|a|$ when all data values are multiplied by a.
- A *shape* parameter is any parameter that is not changed by location and scale parameters.
- Shape parameters dictate skewness and kurtosis.

3.1.16 Location, Scale and Shape Parameters

Examples of location, scale and shape parameters:

- The mean or median of any distribution are location parameters.
- The standard deviation (alternatively, variance) or median absolute deviation of any distribution are scale parameters.
- The degrees of freedom parameter of a t distribution is a shape parameter.

3.2 Heavy-Tailed Distributions

3.2.1 Heavy-Tailed Distributions

Observed data often do not conform to a Normal distribution.

- In many cases, extreme values are more likely than would be dictated by a Normal.
- This is especially true of financial data.
- In this lecture, we will study several examples of distributions with heavy tails, which assign higher probability to extreme values.

3.2.2 Generalized Error Distributions

Suppose that X follows a *Generalized Error Distribution* with shape parameter ν : $X \sim GED(\nu)$.

- Then for $x \in (-\infty, \infty)$,

$$f_X(x|\nu) = \kappa(\nu) \exp\left\{-\frac{1}{2}\left|\frac{x}{\lambda_\nu}\right|^\nu\right\}, \text{where}$$

$$\lambda_\nu = \left(\frac{2^{-2/\nu}\Gamma(\nu^{-1})}{\Gamma(3/\nu)}\right)^{1/2} \quad \text{and} \quad \kappa(\nu) = \frac{\nu}{\lambda_\nu 2^{1+\frac{1}{\nu}}\Gamma(\nu^{-1})}.$$

3.2.3 Generalized Error Distributions

- λ_ν and $\kappa(\nu)$ are constants and are chosen so that the density integrates to unity and has unit variance.
- The shape parameter $\nu > 0$ determines tail weight.

3.2.4 Generalized Error Distributions

For many distributions, the scaling constants are simply a nuisance.

- We can focus attention on only the part of the function that relates to values of the random variable.
- Disregarding constants, we say that the density is *proportional* to:

$$f_X(x|\nu) \propto \exp\left\{-\left|\frac{x}{\theta}\right|^\nu\right\}.$$

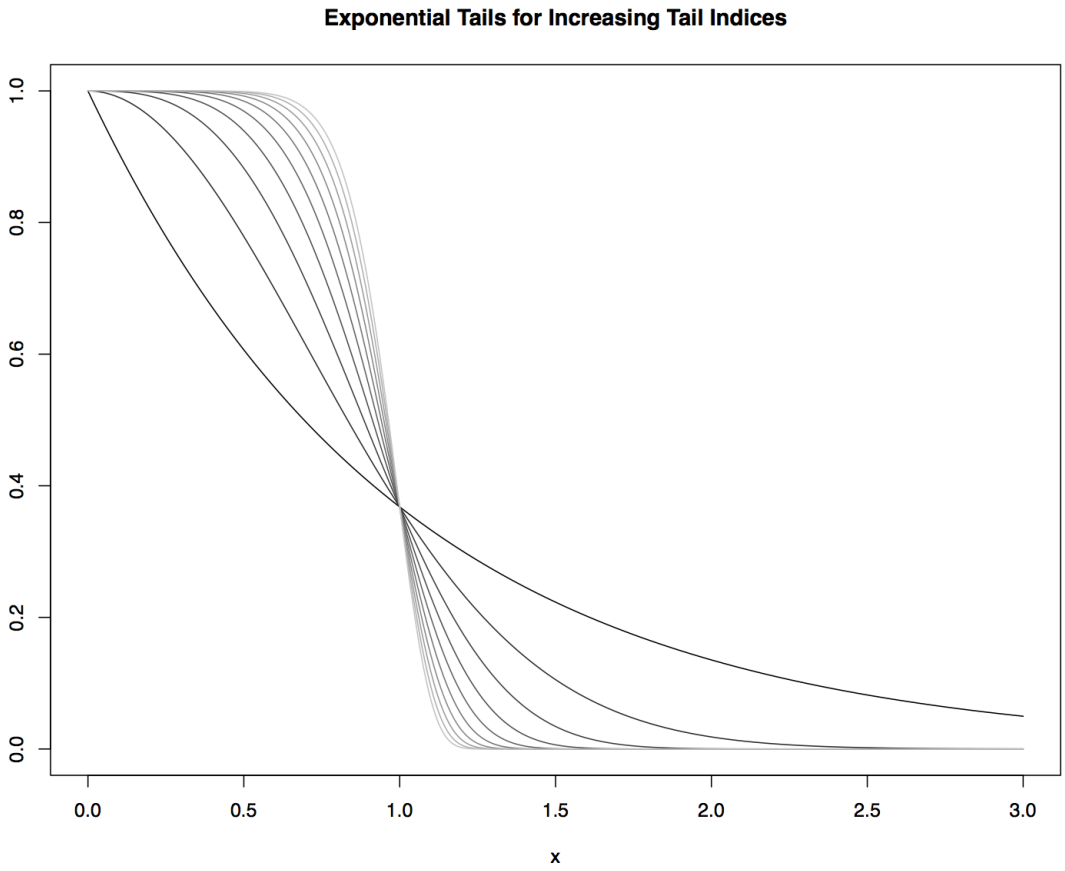
- As $x \to \infty$, $-|x|^\nu \to -\infty$ faster for larger values of ν.
- This means that as $x \to \infty$, $f_X(x|\nu) \to 0$ faster for larger values of ν.

3.2.5 Exponential Tails

For generalized error distributions, larger values of ν correspond to lighter tails and smaller values to heavier tails.

- We say that the generalized error distribution has *exponential tails*, since the tails diminish exponentially as $x \to \infty$ and $x \to -\infty$.

3.2.6 Exponential Tails



To create this plot, run the following script:

```
################################################################################
# Plot Exponential Tails for Varying Tail Indices
################################################################################

# Tail index
alpha = 1:10;

# Grid of values for x variable
xGrid = seq(0,3,length=1000);

# For each alpha, compute exp(-|x|^alpha) and store as column in yGrid matrix
yGrid = NULL;
for(i in 1:length(alpha)){
    yGrid = cbind(yGrid, exp(-abs(xGrid^alpha[i])))
}

# Plot the functions for each alpha
plot(xGrid, yGrid[,1], type='l', xlab='x', ylab='', ylim=c(0,1),
     main='Exponential Tails for Increasing Tail Indices')
for(i in 2:length(alpha)){
    lines(xGrid, yGrid[,i], col=gray(i/13))
}
```

```
dev.copy(pdf, file="expTails.pdf", height=8, width=10)
dev.off()
```

3.2.7 Generalized Error Distribution Examples

Special cases of generalized error distributions:

- $\nu = 2$: $\mathcal{N}(0, 1)$.
- $\nu = 1$: Double-exponential distribution.
- The double-exponential distribution has heavier tails than a standard normal since its shape parameter is smaller.
- Heavier tails that the double-exponential are obtained with $\nu < 1$.

3.2.8 Power-Law Distributions

Suppose that X follows a *Power-Law Distribution* with shape parameter α : $X \sim PL(\alpha)$.

- Then for $x \in (-\infty, \infty)$,

$$f_X(x|\alpha) \propto Ax^{-(1+\alpha)}.$$

- A is chosen so that the density integrates to unity.
- $\alpha > 0$, otherwise the density will integrate to ∞.
- The power-law distribution has a polynomial tail, because the tails diminish at a polynomial rate as $x \to \infty$ and $x \to -\infty$.

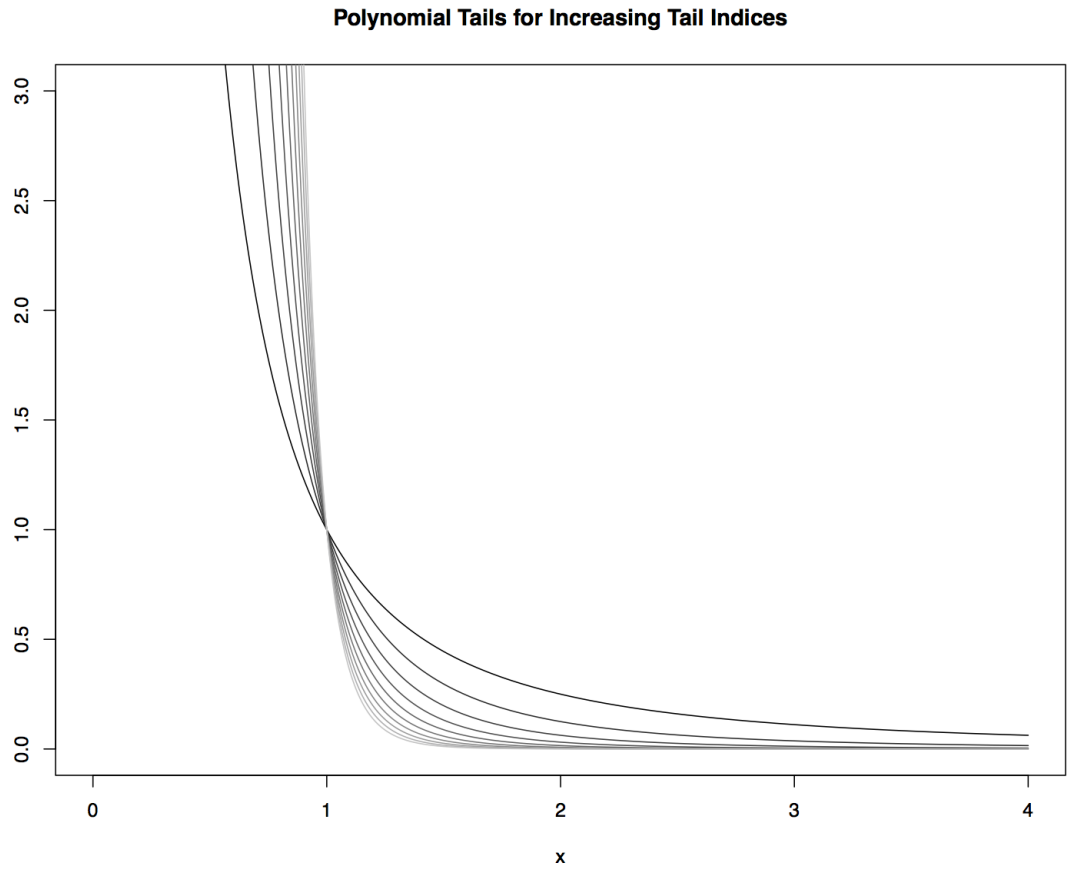
3.2.9 Polynomial Tails

The parameter α is referred to as the *tail index*.

- As $x \to \infty$, $x^{-(1+\alpha)} \to 0$ faster for larger values of α.
- This means that larger values of α correspond to lighter tails and smaller values to heavier tails.
- A power-law distribution has heavier tails than a generalized error distribution:

$$\frac{\exp\left(-\left|\frac{x}{\theta}\right|^{\nu}\right)}{|x|^{-(1+\alpha)}} \to 0 \quad \text{as} \quad |x| \to \infty.$$

3.2.10 Polynomial Tails



To create this plot, run the following script:

```
################################################################################
# Plot Polynomial Tails for Varying Tail Indices
################################################################################

# Tail index
alpha = 1:10;

# Grid of values for x variable
xGrid = seq(0,4,length=1000);

# For each alpha, compute x^(-(1+alpha)) and store as column in yGrid matrix
yGrid = NULL;
for(i in 1:length(alpha)){
    yGrid = cbind(yGrid, xGrid^(-(1+alpha[i])))
}

# Plot the functions for each alpha
plot(xGrid, yGrid[,1], type='l', xlab='x', ylab='',
     main='Polynomial Tails for Increasing Tail Indices', ylim=c(0,3))
for(i in 2:length(alpha)){
    lines(xGrid, yGrid[,i], col=gray(i/13))
}
```

```
dev.copy(pdf, file="polyTails.pdf", height=8, width=10)
dev.off()
```

3.2.11 t-Distribution

The density of a t -distribution is

$$f_{t,\nu}(y) = \left[\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\sqrt{\pi\nu}\Gamma\left(\frac{\nu}{2}\right)}\right] \frac{1}{(1+y^2/\nu)^{\frac{\nu+1}{2}}},$$

where

$$\Gamma(t) = \int_0^\infty x^{t-1}\exp(-x)dx, \quad t > 0.$$

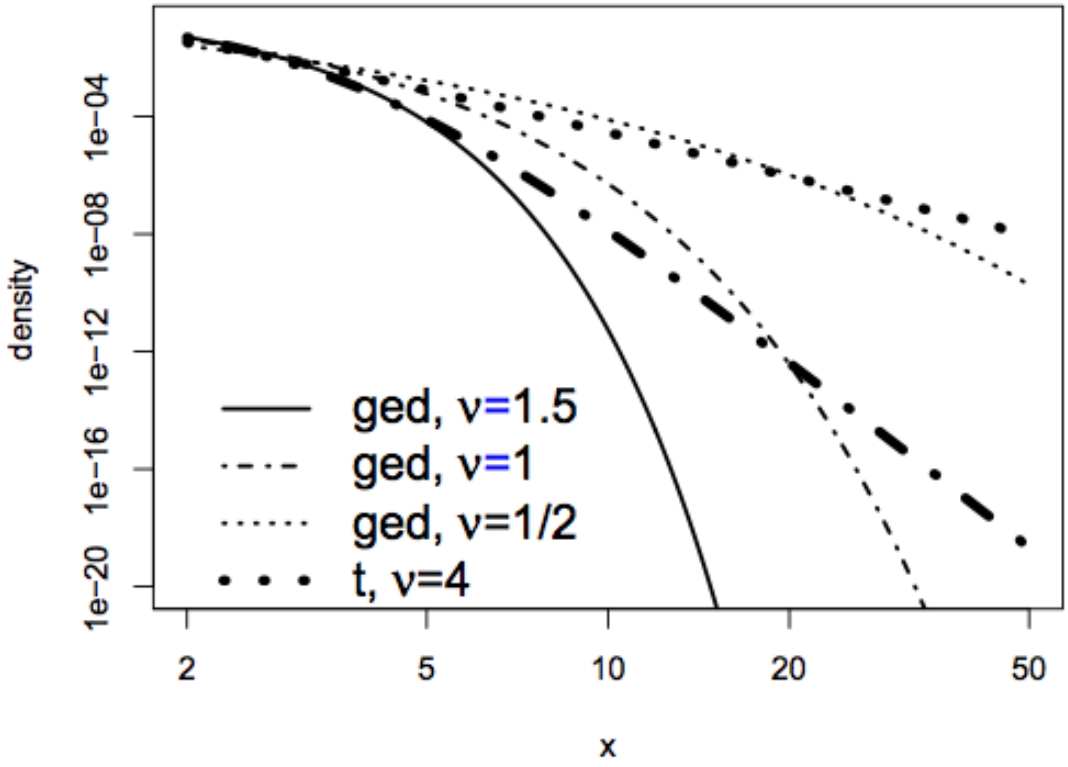
3.2.12 t -Distribution

Note that for large values of $|y|$,

$$f_{t,\nu}(y) \propto \frac{1}{(1+y^2/\nu)^{\frac{\nu+1}{2}}} \tilde{\propto} \frac{1}{(y^2/\nu)^{\frac{\nu+1}{2}}} \propto |y|^{-(\nu+1)}.$$

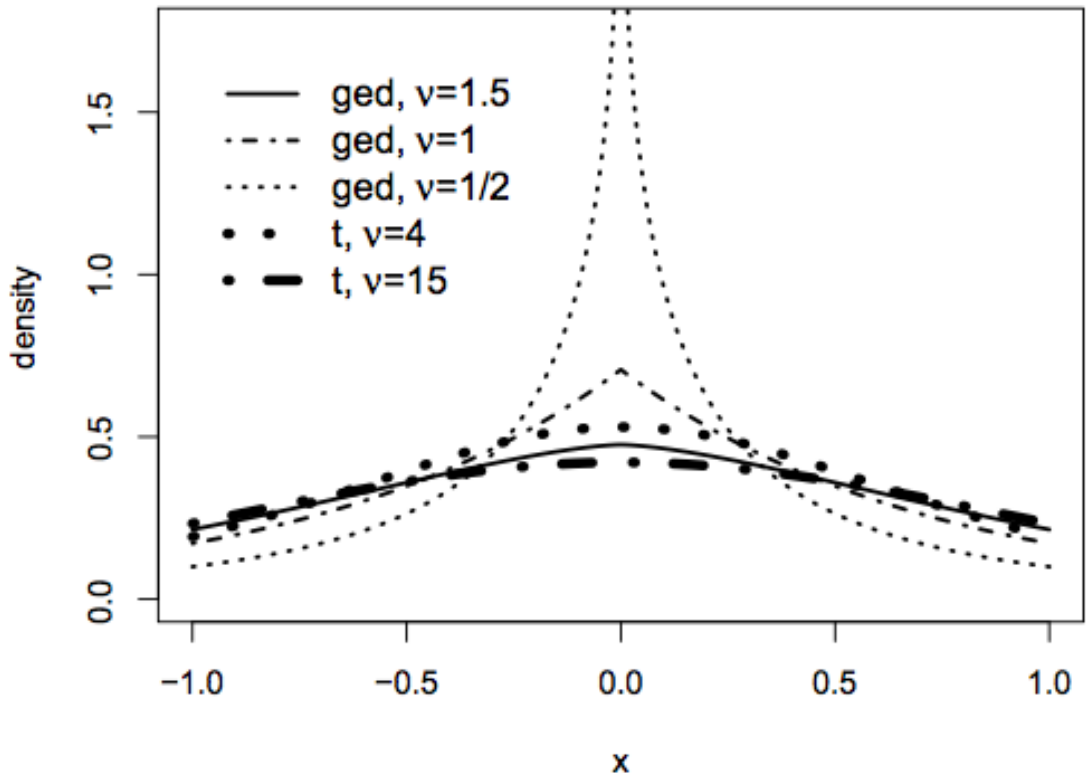
- This means the t-distribution has polynomial tails with tail index ν.
- Smaller values of ν correspond to heavier tails.

3.2.13 Comparison of Gen. Error and t- Dist



This plot was taken directly from Ruppert (2011).

3.2.14 Comparison of Gen. Error and t- Dist



This plot was taken directly from Ruppert (2011).

3.2.15 Discrete Mixtures

Consider a distribution that is 90% $\mathcal{N}(0, 1)$ and 10% $\mathcal{N}(0, 25)$.

- Generate $X \sim \mathcal{N}(0, 1)$.
- Generate $U \sim Unif(0, 1)$, with U independent of X.
- Set $Y = X$ if $U < 0.9$.
- Set $Y = 5X$ if $U \geq 0.9$.

3.2.16 Discrete Mixtures

We say that Y follows a *finite* or *discrete normal mixture distribution*.

- Roughly 90% of the time it is drawn from a $\mathcal{N}(0, 1)$.
- Roughly 10% of the time it is drawn from a $\mathcal{N}(0, 25)$.
- The individual normal distributions are called the *component* distributions of Y.
- This random variable could be used to model a market with two *regimes*: low volatility and high volatility.

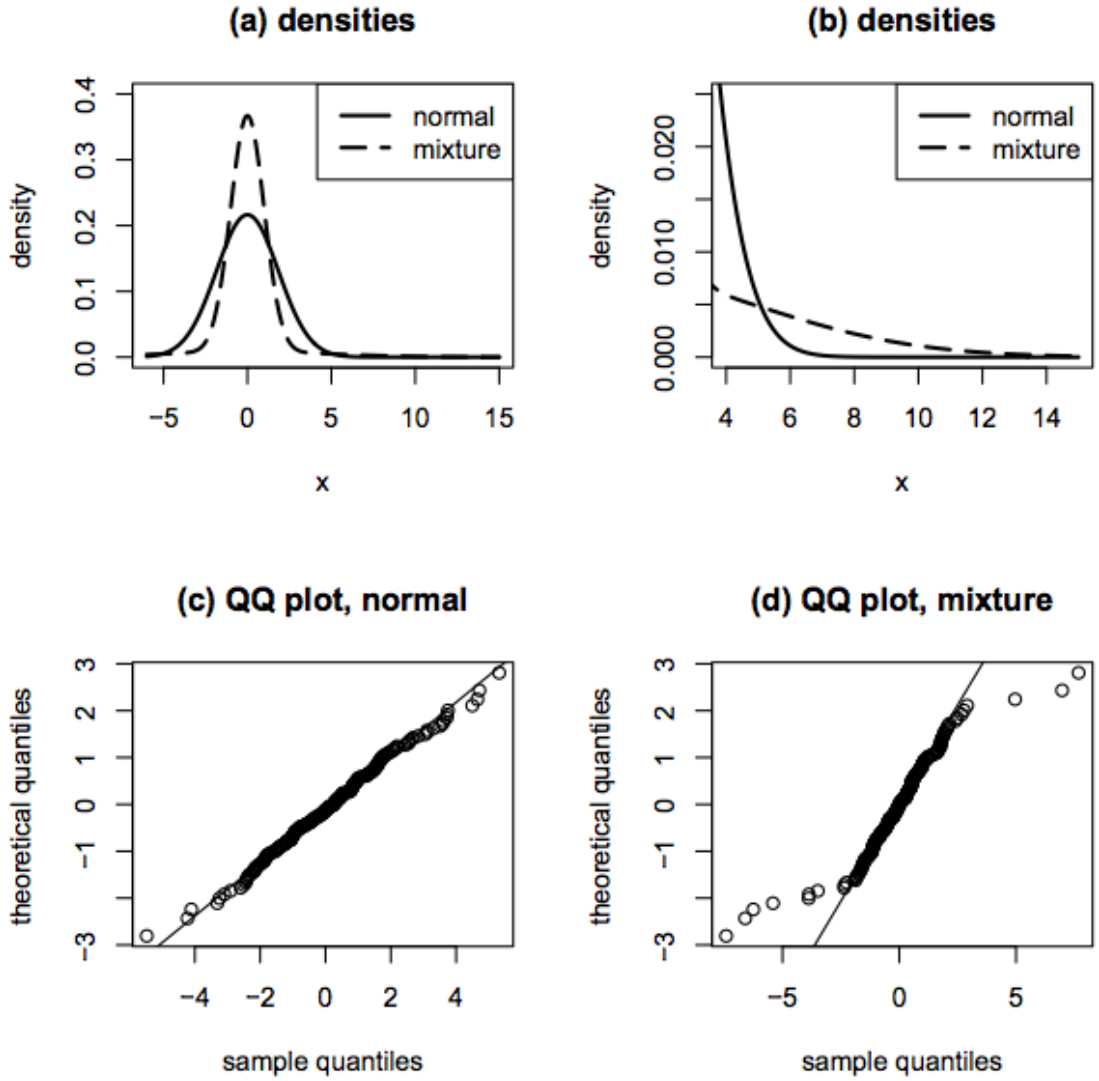
3.2.17 Discrete Mixtures

The variance of Y is

$$Var(Y) = 0.9 \times 1 + 0.1 \times 25 = 3.4.$$

- Consider $Z \sim \mathcal{N}(0, \sqrt{3.4}) = \mathcal{N}(0, 1.84)$.
- The distributions of Y and Z are *very* different.
- f_Y has much heavier tails than f_Z.
- For example, the probability of observing the value 6 (3.25 standard deviations from the mean) is essentially zero for Z.
- However, 10% of the time, the value 6 is only 6/5 = 1.2 standard deviations from the mean for Y.

3.2.18 Discrete Mixtures



This plot was taken directly from Ruppert (2011).

3.2.19 Continuous Mixtures

In general, Y follows a *normal scale mixture distribution* if

$$Y = \mu + \sqrt{U}Z,$$

where

- μ is a constant.
- $Z \sim \mathcal{N}(0, 1)$.
- U is a positive random variable giving the variance of each normal component.
- Z and U are independent.

3.2.20 Continuous Mixtures

- If U is continuous, Y follows a *continuous scale mixture distribution*.
- f_U is known as the *mixing distribution*.
- A finite normal mixture has exponential tails.
- A continuous normal mixture can have polynomial tails if the mixing distribution has heavy enough tails.
- The t -distribution is an example of a continuous normal mixture.

3.3 Maximum Likelihood Estimation

3.3.1 Estimating Parameters of Distributions

We almost never know the true distribution of a data sample.

- We might hypothesize a family of distributions that capture broad characteristics of the data (locations, scale and shape).
- However, there may be a set of one or more parameters of the distribution that we don't know.
- Typically we use the data to estimate the unknown parameters.

3.3.2 Joint Densities

Suppose we have a collection of random variables $\mathbf{Y} = (Y_1, \ldots, Y_n)'$.

- We view a data sample of size n as one realization of each random variable: $\mathbf{y} = (y_1, \ldots, y_n)'$.
- The *joint cumulative density* of $\mathbf{Y}$ is

$$F_{\mathbf{Y}}(\mathbf{y}) = P(Y_1 \leq y_1, \ldots, Y_n \leq y_n).$$

3.3.3 Joint Densities

- The *joint probability density* of $\mathbf{Y}$ is

$$f_{\mathbf{Y}}(\mathbf{y}) = \frac{\partial^n F_{\mathbf{Y}}(\mathbf{y})}{\partial Y_1 \dots \partial Y_n}.$$

since

$$F_{\mathbf{Y}}(\mathbf{y}) = \int_{-\infty}^{y_1} \dots \int_{-\infty}^{y_n} f_{\mathbf{Y}}(\mathbf{a})\ da_1 \dots da_n.$$

3.3.4 Independence

When $Y_1, \dots, Y_n$ are independent of each other and have identical distributions:

- We say that they are *independent and identically distributed*, or i.i.d.
- When $Y_1, \dots, Y_n$ are i.i.d., they have the same marginal densities:

$$f_{Y_1}(y) = \dots = f_{Y_n}(y).$$

3.3.5 Independence

Further, when $Y_1, \dots, Y_n$ are i.i.d.

$$f_{\mathbf{Y}}(\mathbf{y}) = f_{Y_1}(y_1) \cdot f_{Y_2}(y_2) \cdots f_{Y_n}(y_n) = \prod_{i=1}^{n} f_{Y_i}(y_i).$$

- This is analogous to the computation of joint probabilities.
- For independent events A, B and C,

$$P(A \cap B \cap C) = P(A)P(B)P(C).$$

3.3.6 Maximum Likelihood Estimation

One of the most important and powerful methods of parameter estimation is *maximum likelihood estimation*. It requires

- A data sample: $\mathbf{y} = (y_1, \dots, y_n)'$.
- A joint probability density:

$$f_{\mathbf{Y}}(\mathbf{y}|\theta) = \prod_{i=1}^{n} f_{Y_i}(y_i|\theta).$$

where θ is a vector of parameters.

3.3.7 Likelihood

$f_{\mathbf{Y}}(\mathbf{y}|\theta)$ is loosely interpreted as the probability of observing data sample $\mathbf{y}$, given a functional form for the density of $Y_1, \ldots, Y_n$ and given a set of parameters θ.

- We can reverse the notion and think of $\mathbf{y}$ as being fixed and θ some unknown variable.
- In this case we write $\mathcal{L}(\theta|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}|\theta)$.
- We refer to $\mathcal{L}(\theta|\mathbf{y})$ as the likelihood.
- Fixing $\mathbf{y}$, maximum likelihood estimation chooses the value of θ that maximizes $\mathcal{L}(\theta|\mathbf{y}) = f_{\mathbf{Y}}(\mathbf{y}|\theta)$.

3.3.8 Likelihood Maximization

Given $\theta = (\theta_1, \ldots, \theta_p)'$, we maximize $\mathcal{L}(\theta|\mathbf{y})$ by

- Differentiating with respect to each θ_i, $i = 1, \ldots, p$.
- Setting the resulting derivatives equal to zero.
- Solving for the values $\hat{\theta}_i$, $i = 1, \ldots, p$, that make all of the derivatives zero.

3.3.9 Log Likelihood

It is often easier to work with the logarithm of the likelihood function.

- By the properties of logarithms

$$\begin{aligned}
\ell(\theta|\mathbf{y}) &= \log\left(\mathcal{L}(\theta|\mathbf{y})\right) \\
&= \log\left(f_{\mathbf{Y}}(\mathbf{y}|\theta)\right) \\
&= \log\left(\prod_{i=1}^{n} f_{Y_i}(y_i|\theta)\right) \\
&= \sum_{i=1}^{n} \log\left(f_{Y_i}(y_i|\theta)\right).
\end{aligned}$$

3.3.10 Log Likelihood

- Maximizing $\ell(\theta|\mathbf{y})$ is the same as maximizing $\mathcal{L}(\theta|\mathbf{y})$ since $\log$ is a monotonic transformation.
- A derivative of $\mathcal{L}$ will involve many chain-rule products, whereas a derivative of ℓ will simply be a sum of derivatives.

3.3.11 MLE Example

Suppose we have a dataset $\mathbf{y} = (y_1, \ldots, y_n)$, where $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$.

- We will assume μ is *unknown* and σ is *known*.
- So, $\theta = \mu$ (it is a single value, rather than a vector).

3.3.12 MLE Example

- The likelihood is

$$\begin{aligned}\mathcal{L}(\mu|\mathbf{y}) &= f_{\mathbf{Y}}(\mathbf{y}|\mu) \\ &= \prod_{i=1}^{n} f_{Y_i}(y_i|\mu) \\ &= \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2}\frac{(y_i-\mu)^2}{\sigma^2}\right\} \\ &= \frac{1}{(2\pi\sigma^2)^{n/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2\right\}.\end{aligned}$$

3.3.13 MLE Example

The log likelihood is

$$\ell(\mu|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i-\mu)^2.$$

3.3.14 MLE Example

- The MLE, $\hat{\mu}$, is the value that sets $\frac{d}{d\mu}\ell(\mu|\mathbf{y}) = 0$:

$$\begin{aligned}\frac{d}{d\mu}\ell(\mu|\mathbf{y})\bigg|_{\hat{\mu}} &= \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i-\hat{\mu}) = 0 \\ \Rightarrow \sum_{i=1}^{n}(y_i-\hat{\mu}) &= 0 \\ \Rightarrow \sum_{i=1}^{n}\hat{\mu} &= \sum_{i=1}^{n} y_i \\ \Rightarrow \hat{\mu} = \bar{y} &= \frac{1}{n}\sum_{i=1}^{n} y_i.\end{aligned}$$

3.3.15 MLE Example: $n = 1$, Unknown μ

Suppose we have only one observation: y_1.

- If we specialize the previous result:

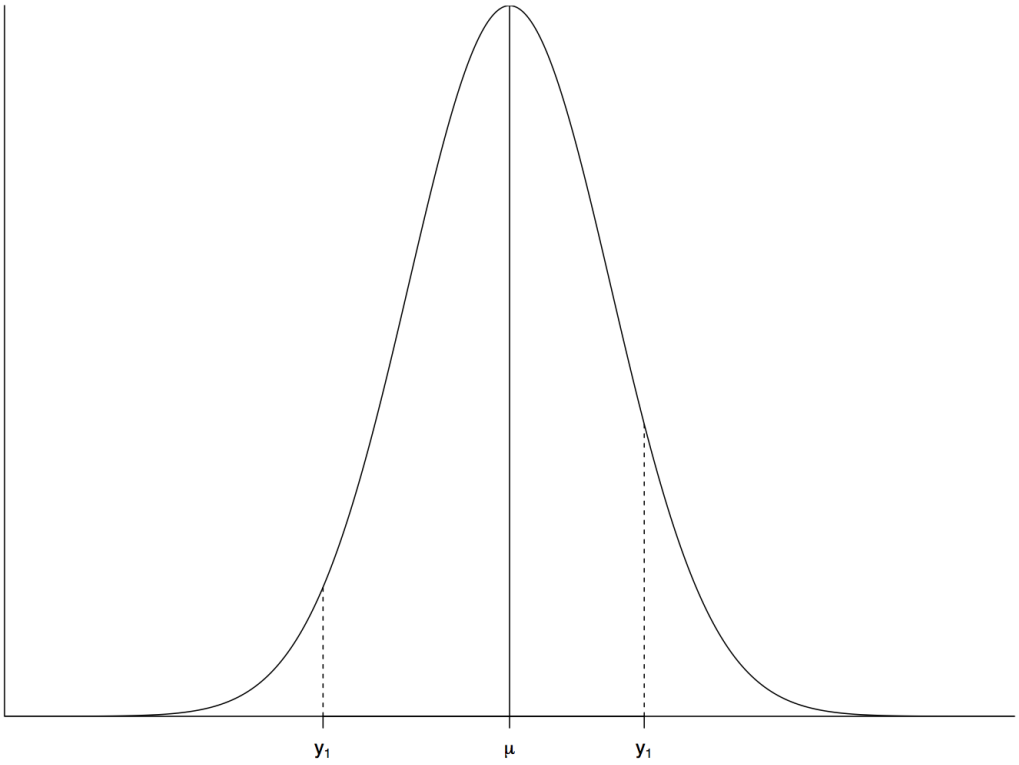
$$\hat{\mu} = y_1.$$

- The density $f_{Y_1}(y_1|\mu)$ gives the probability of observing some data value y_1, conditional on some *known* parameter μ.
- This is a normal distribution with mean μ and variance σ^2.

3.3.16 MLE Example: $n = 1$, Unknown μ

- The likelihood $\mathcal{L}(\mu|y_1)$ gives the probability of μ, conditional on some observed data value y_1.
- This is a normal distribution with mean y_1 and variance σ^2.

3.3.17 MLE Example: $n = 1$

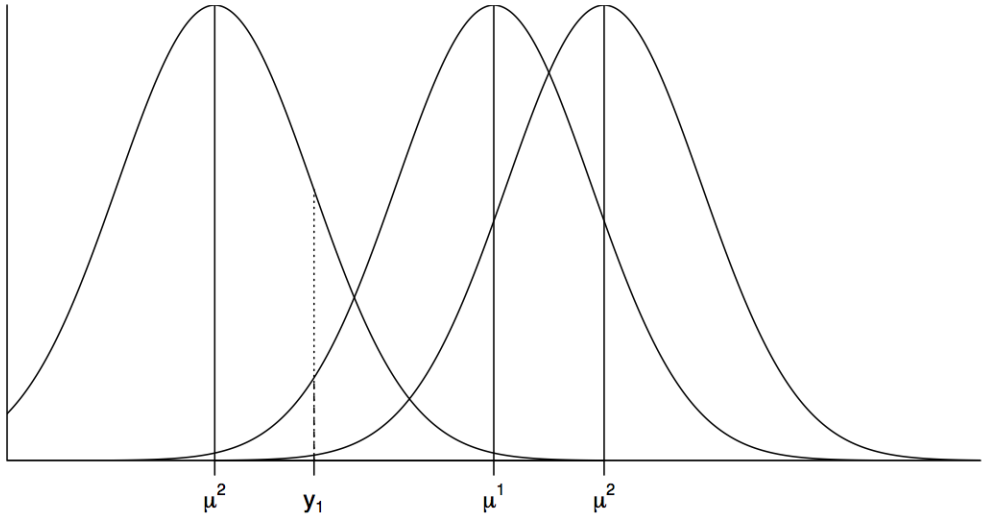


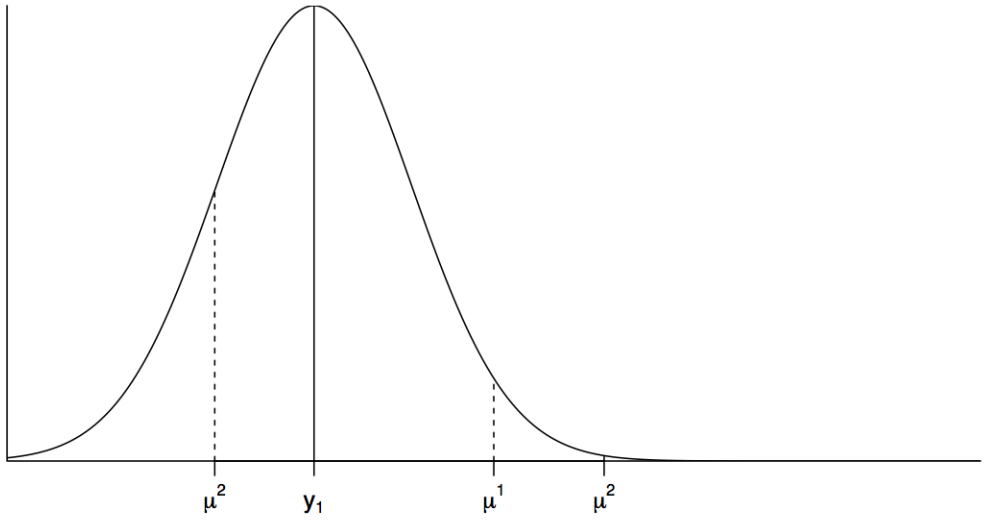
To create this plot, run the following script:

```
################################################################################
# Plot likelihood for normal data with unknown mean
################################################################################

# Generate the true normal density
mu = 10;
sig = 15;
xGrid = seq(mu-5*sig, mu+5*sig, length=10000)
trueDens = dnorm(xGrid, mu, sig)

# A couple of possible data values
y11 = 30;
y12 = -17.7;

# Plot the true normal distribution with possible data values
```

```
plot(xGrid, trueDens, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(mu), expression(y[1]), expression(y[1])),
     at=c(mu, y11, y12))
abline(v=mu)
segments(y11, 0, y11, dnorm(y11, mu, sig), lty=2)
segments(y12, 0, y12, dnorm(y12, mu, sig), lty=2)
dev.copy(pdf, file="densExample.pdf", height=8, width=10)
dev.off()
```

3.3.18 MLE Example: $n = 1$, Unknown μ





To create this plot, run the following script:

```
# Plot several densities for fixed data observation
mu1 = -33
mu2 = 27
dens1 = dnorm(xGrid, mu1, sig)
```

```
dens2 = dnorm(xGrid, mu2, sig)
par(mfrow=c(2,1))
plot(xGrid, trueDens, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(mu^1), expression(mu^2), expression(mu^2), expression(y[1])),
     at=c(mu, mu1, mu2, y12))
lines(xGrid, dens1)
lines(xGrid, dens2)
abline(v=mu)
abline(v=mu1)
abline(v=mu2)
segments(y12, 0, y12, dnorm(y12, mu, sig), lty=2)
segments(y12, 0, y12, dnorm(y12, mu1, sig), lty=3)
segments(y12, 0, y12, dnorm(y12, mu2, sig), lty=4)

# Plot the resulting likelihood
like = dnorm(xGrid, y12, sig)
plot(xGrid, like, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(mu^1), expression(mu^2), expression(mu^2), expression(y[1])),
     at=c(mu, mu1, mu2, y12))
abline(v=y12)
segments(mu, 0, mu, dnorm(mu, y12, sig), lty=2)
segments(mu1, 0, mu1, dnorm(mu1, y12, sig), lty=2)
segments(mu2, 0, mu2, dnorm(mu2, y12, sig), lty=2)
dev.copy(pdf, file="likeExample.pdf", height=10, width=8)
dev.off()
```

3.3.19 MLE Example: $n = 1$, Unknown σ

Let's continue with the assumption of one data observation, y_1.

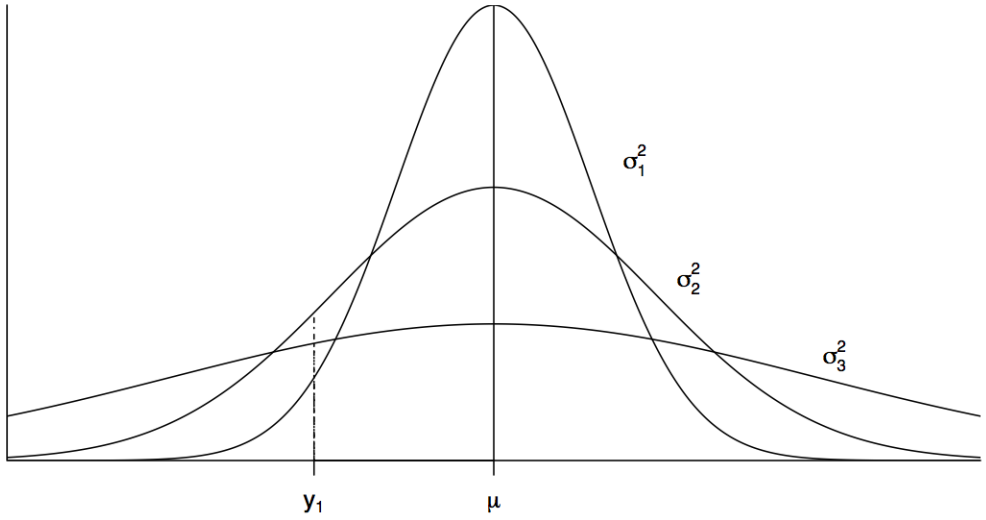
- If μ is known but σ is unknown, the density of the data, y_1, is still normal.
- However, the likelihood is

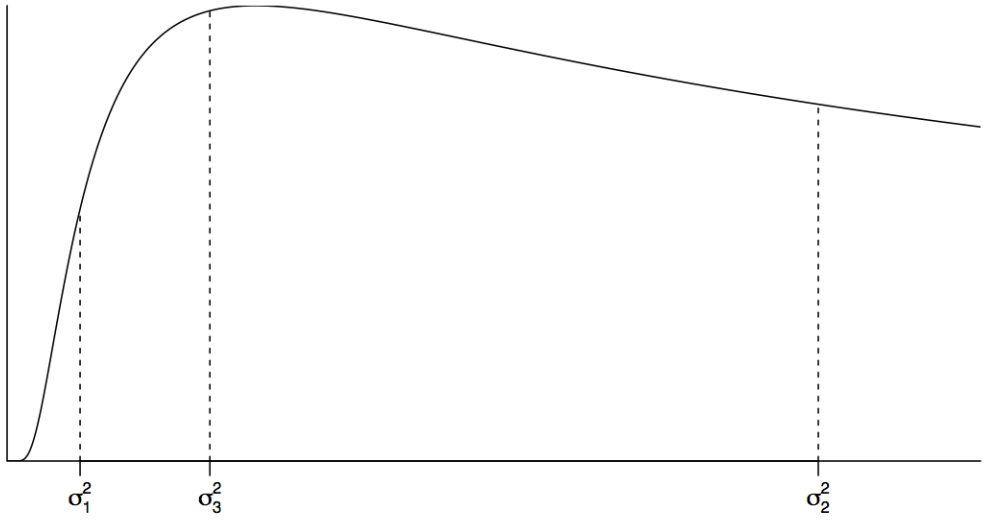
$$\mathcal{L}(\sigma^2|y_1) = \frac{\alpha}{\sigma^2}\exp\left\{-\frac{\beta}{\sigma^2}\right\}$$

$$\alpha = \frac{1}{\sqrt{2\pi}}, \qquad \beta = \frac{(y_1-\mu)^2}{2}.$$

- The likelihood for σ^2 is *not* normal, but *inverse gamma*.

3.3.20 MLE Example: $n = 1$, Unknown σ





To create this plot, run the following script:

```
##############################################################################
# Plot likelihood for normal data with unknown sd
##############################################################################
```

```
# Plot several densities for fixed data observation
sig1 = 50
sig2 = 25
dens1 = dnorm(xGrid, mu, sig1)
dens2 = dnorm(xGrid, mu, sig2)
par(mfrow=c(2,1))
yMax = max(max(trueDens), max(dens1), max(dens2))
plot(xGrid, trueDens, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L', ylim=c(0,yMax))
axis(1, labels=c(expression(mu), expression(y[1])), at=c(mu, y12))
lines(xGrid, dens1)
lines(xGrid, dens2)
abline(v=mu)
segments(y12, 0, y12, dnorm(y12, mu, sig), lty=2)
segments(y12, 0, y12, dnorm(y12, mu, sig1), lty=3)
segments(y12, 0, y12, dnorm(y12, mu, sig2), lty=4)
xDist = max(xGrid)-mu
text(c(mu+0.29*xDist, mu+0.4*xDist, mu+0.7*xDist), c(0.66, 0.4, 0.23)*yMax,
     labels=c(expression(sigma[1]^2), expression(sigma[2]^2), expression(sigma[3]^2)))

# Plot the resulting likelihood (which is an inverse gamma)
alpha = -0.5
beta = ((y12 - mu)^2)/2
scale = 1/sqrt(2*pi)
sigGrid = seq(0.01,3000,length=10000)
like = scale*(sigGrid^(-alpha-1))*exp(-beta/sigGrid)
likeTrue = scale*(sig^(-2*alpha-2))*exp(-beta/sig^2)
like1 = scale*(sig1^(-2*alpha-2))*exp(-beta/sig1^2)
like2 = scale*(sig2^(-2*alpha-2))*exp(-beta/sig2^2)
plot(sigGrid, like, type='l', xaxs='i', yaxs='i', xaxt='n', xlab='', ylab='',
     yaxt='n', bty='L')
axis(1, labels=c(expression(sigma[1]^2), expression(sigma[2]^2), expression(sigma[3]^2)),
     at=c(sig^2, sig1^2, sig2^2))
segments(sig^2, min(like), sig^2, likeTrue, lty=2)
segments(sig1^2, min(like), sig1^2, like1, lty=2)
segments(sig2^2, min(like), sig2^2, like2, lty=2)
dev.copy(pdf, file="likeExample2.pdf", height=10, width=8)
dev.off()
```

3.3.21 MLE Accuracy

Maximum likelihood estimation results in estimates of true unknown parameters.

- What is the probability that our estimates are identical to the true population parameters?
- Our estimates are imprecise and contain error.
- We would like to quantify the precision of our estimates with standard errors.
- We will use the *Fisher Information* to compute standard errors.

3.3.22 Fisher Information

Suppose our likelihood is a function of a single parameter, θ: $\mathcal{L}(\theta|\mathbf{y})$.

- The Fisher Information is

$$\mathcal{I}(\theta) = -E\left[\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{y})\right].$$

- The observed Fisher Information is

$$\widetilde{\mathcal{I}}(\theta) = -\frac{d^2}{d\theta^2}\ell(\theta|\mathbf{y}).$$

3.3.23 Fisher Information

- Observed Fisher Information does not take an expectation, which may be difficult to compute.
- Since $\ell(\theta|\mathbf{y})$ is often a sum of many terms, $\widetilde{\mathcal{I}}(\theta)$ will converge to $\mathcal{I}(\theta)$ for large samples.

3.3.24 MLE Central Limit Theorem

Under certain conditions, a central limit theorem holds for the MLE, $\hat{\theta}$.

- For infinitely large samples $\mathbf{y}$,

$$\hat{\theta} \sim \mathcal{N}(\theta, \mathcal{I}(\theta)^{-1}).$$

- For large samples, $\hat{\theta}$ is normally distributed *regardless* of the distribution of the data, $\mathbf{y}$.

3.3.25 MLE Central Limit Theorem

- $\hat{\theta}$ is also normally distributed for large samples even if $\mathcal{L}(\theta|\mathbf{y})$ is some other distribution.
- Hence, for large samples,

$$Var(\hat{\theta}) = \frac{1}{\mathcal{I}(\theta)} \qquad \Rightarrow \qquad Std(\hat{\theta}) = \frac{1}{\sqrt{\mathcal{I}(\theta)}}.$$

3.3.26 MLE Standard Errors

Since we don't know the true θ, we approximate

$$Std(\hat{\theta}) \approx \frac{1}{\sqrt{\mathcal{I}(\hat{\theta})}}.$$

- Alternatively, to avoid computing the expectation, we could use the approximation

$$Std(\hat{\theta}) \approx \frac{1}{\sqrt{\widetilde{\mathcal{I}}(\hat{\theta})}}.$$

3.3.27 MLE Standard Errors

- In reality, we never have an infinite sample size.
- For finite samples, these values are approximations of the standard error of $\hat{\theta}$.

3.3.28 MLE Variance Example

Let's return to the example where $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, with known σ.

- The log likelihood is

$$\ell(\mu|\mathbf{y}) = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log(\sigma^2) - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(y_i - \mu)^2.$$

- The resulting derivatives are

$$\frac{\partial \ell(\mu|\mathbf{y})}{\partial \mu} = \frac{1}{\sigma^2}\sum_{i=1}^{n}(y_i - \mu), \qquad \frac{\partial^2 \ell(\mu|\mathbf{y})}{\partial \mu^2} = -\frac{n}{\sigma^2}.$$

3.3.29 MLE Variance Example

In this case the Fisher Information is identical to the observed Fisher Information:

$$\mathcal{I}(\mu) = -E\left[-\frac{n}{\sigma^2}\right] = \frac{n}{\sigma^2} = \widetilde{\mathcal{I}}(\mu).$$

- Since $\mathcal{I}(\mu)$ doesn't depend on μ, we don't need to resort to an approximation with $\hat{\mu} = \bar{\mathbf{y}}$.
- The result is

$$Std(\hat{\mu}) = \frac{1}{\sqrt{\mathcal{I}(\mu)}} = \frac{\sigma}{\sqrt{n}}.$$

CHAPTER

FOUR

RESAMPLING

4.1 Properties of Estimators

We've now seen that estimation is not enough.

- We often want to know about properties of estimators.
- For example, what is the standard error of an estimator?
- Remember that before data is observed, an estimator is a random variable itself.
- As an example, $\bar{Y}$ is a sum of random variables divided by a constant value.
- Before $\{Y_i\}_{i=1}^n$ is observed, $\bar{Y}$ is random and has its own variance and standard deviation.

4.2 Resampling

It is often challenging or impossible to compute certain characteristics of estimators.

- We would like to replace theoretical calculations with Monte Carlo simulation, which draws additional samples from the population.
- Sampling from the true population is typically impossible.

4.3 Resampling

We substitute sampling from the true population with sampling from the observed sample.

- This is referred to as *resampling*.
- If the sample is a good representation of the true population, then sampling from the sample should approximate sampling from the population.

4.4 Bootstrapping

Suppose the original sample has n data observations.

- *Bootstrapping* involves drawing B new samples of size n from the original sample.
- Each bootstrap sample is done with replacement.
- Otherwise, the bootstrap samples would all be identical to the original sample (why?).

- Drawing with replacement allows each bootstrap observation to be drawn in an $i.i.d.$ fashion from the sample.
- So, the original sample plays the role of the population.

4.5 Bootstrap Estimates

Let θ be a parameter of interest and let $\hat{\theta}$ denote an estimate of θ using a sample of data, $\{y_i\}_{i=1}^n$.

- $\hat{\theta}$ might be calculated by maximum likelihood estimation.
- We could create B new samples from $\{y_i\}_{i=1}^n$ by resampling with replacement.
- For each new sample $j = 1, \ldots, B$, we could compute $\hat{\theta}_j^*$ in the exact way $\hat{\theta}$ was computed with $\{y_i\}_{i=1}^n$.

4.6 Bootstrap Estimates

- One way to estimate $E[\hat{\theta}]$ is by averaging the bootstrap estimates:

$$E[\hat{\theta}] \approx \bar{\hat{\theta}}^* = \frac{1}{B} \sum_{j=1}^{B} \hat{\theta}_j^*.$$

4.7 Estimating Bias

True bias for an estimator is defined as

$$\text{BIAS}(\hat{\theta}) = E[\hat{\theta}] - \theta.$$

- We can approximate the population average, $E[\hat{\theta}]$, with a bootstrap average, $\bar{\hat{\theta}}^*$:

$$\text{BIAS}_{\text{boot}}(\hat{\theta}) = \bar{\hat{\theta}}^* - \hat{\theta}.$$

- We replaced the true population value, θ, with the sample value, $\hat{\theta}$, since the sample substitutes for the population.

4.8 Estimating Standard Error

The true standard deviation of $\hat{\theta}$ can be estimated with the bootstrap estimates:

$$s_{\text{boot}}(\hat{\theta}) = \sqrt{\frac{1}{B-1} \sum_{j=1}^{B} \left(\hat{\theta}_j^* - \bar{\hat{\theta}}^*\right)^2}.$$

4.9 Example: Pareto Distribution

Suppose we have a sample of random variables drawn from a Pareto distribution:

$$Y_i \overset{i.i.d.}{\sim} \mathcal{P}(\alpha, \beta), \qquad i = 1, \ldots, n.$$

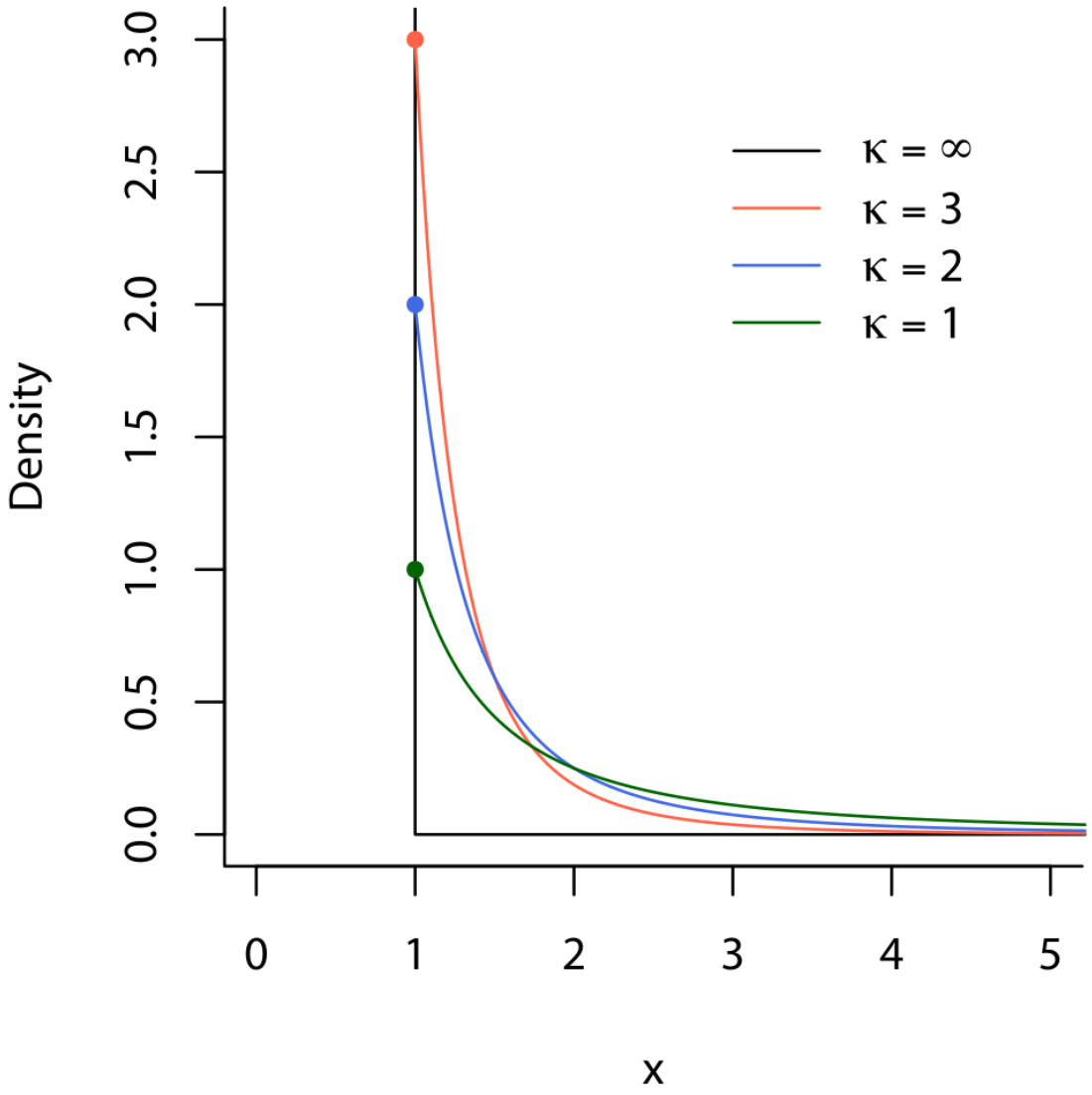
- The density of each Y_i is

$$f(y|\alpha, \beta) = \frac{\beta \alpha^{\beta}}{y^{\beta+1}}.$$

- If $Y_i \sim \mathcal{P}(\alpha, \beta)$, then $\alpha > 0$, $\beta > 0$ and $Y_i > \alpha$.
- α is a parameter dictating the minimum possible value of Y_i and β is a shape parameter.

4.10 Example: Pareto Distribution

In the graph below, $\beta = \kappa$.



4.11 Example: Pareto Distribution

The joint density of $\mathbf{Y} = (Y_1, \ldots, Y_n)'$ is

$$f_{\mathbf{Y}}(\mathbf{y}|\alpha, \beta) = \prod_{i=1}^{n} f_{Y_i}(y_i|\alpha, \beta)$$

$$= \prod_{i=1}^{n} \frac{\beta \alpha^{\beta}}{y_i^{\beta+1}}$$

$$= \frac{\beta^n \alpha^{n\beta}}{\prod_{i=1}^{n} y_i^{\beta+1}}.$$

4.12 Example: Pareto Distribution

Assuming α is known, the log likelihood of β is

$$\ell(\beta|\alpha, \mathbf{y}) = \log(f_{\mathbf{Y}}(\mathbf{y}|\alpha, \beta))$$

$$= n\log(\beta) + n\beta\log(\alpha) - (1+\beta)\sum_{i=1}^{n}\log(y_i).$$

4.13 Example: Pareto Distribution

The MLE, $\hat{\beta}$ is the value such that

$$\left.\frac{\partial \ell}{\partial \beta}\right|_{\beta=\hat{\beta}} = \frac{n}{\hat{\beta}} + n\log(\alpha) - \sum_{i=1}^{n}\log(y_i) = 0.$$

$$\Rightarrow \hat{\beta} = \frac{n}{\sum_{i=1}^{n}\log(y_i) - n\log(\alpha)}.$$

4.14 Example: Pareto Distribution

The second derivative of the log likelihood is

$$\frac{\partial^2 \ell}{\partial \beta^2} = -\frac{n}{\beta^2}.$$

- The observed Fisher information is

$$\tilde{\mathcal{I}}(\hat{\beta}) = -\left.\frac{\partial^2 \ell}{\partial \beta^2}\right|_{\beta=\hat{\beta}} = \frac{n}{\hat{\beta}^2}.$$

- The asymptotic standard error of $\hat{\beta}$ is

$$Std(\hat{\beta}) \approx \sqrt{\tilde{\mathcal{I}}(\hat{\beta})^{-1}} = \frac{\hat{\beta}}{\sqrt{n}}.$$

4.15 Example: Pareto Distribution

Given a sample of n observations from a Pareto distribution:

- We can compute the MLE, $\hat{\beta}$.
- We can compute the asymptotic standard error $\hat{\beta}/\sqrt{n}$.

4.16 Example: Pareto Distribution

We can generate B new samples by resampling.

- For each new sample, we can compute $\hat{\beta}_j$, $j = 1, \ldots, B$.
- We can compute the standard deviation of $\{\hat{\beta}_j\}_{j=1}^B$ and compare to the asymptotic standard error.
- The bootstrap standard error will be a better estimate of variation than the asymptotic standard error when n is small.

4.17 Bootstrap Confidence Intervals

Given a set of bootstrap estimates, $\{\hat{\theta}_j^*\}_{j=1}^B$, we can form a $1-\alpha$ confidence interval with the normal approximation

$$\left(\hat{\theta} - s_{\text{boot}}(\hat{\theta})\, z_{\alpha/2},\ \hat{\theta} + s_{\text{boot}}(\hat{\theta})\, z_{\alpha/2}\right)$$

where $z_{\alpha/2}$ is the α -upper quantile of the standard normal distribution.

- Note that the interval is centered around $\hat{\theta}$ rather than θ.
- In this case $\hat{\theta}$ is substituted for θ, just as the data sample is substituted for the true population.

4.18 Bootstrap Confidence Intervals

Alternatively, we could compute the α and $1-\alpha$ empirical quantiles of the bootstrap estimates, $\{\hat{\theta}_j^*\}_{j=1}^B$: $q_{\alpha/2}$ and $q_{(1-\alpha)/2}$.

- The resulting $1-\alpha$ confidence interval is

$$\left(q_{\alpha/2}, q_{(1-\alpha)/2}\right).$$

CHAPTER

FIVE

TIME SERIES

Contents:

5.1 Stationarity

5.1.1 Time Series

A time series is a stochastic process indexed by time:

$$Y_1, Y_2, Y_3, \ldots, Y_{T-1}, Y_T.$$

- *Stochastic* is a synonym for *random*.
- So a time series is a sequence of (potentially different) random variables ordered by time.
- We will let lower-case letters denote a realization of a time series.

$$y_1, y_2, y_3, \ldots, y_{T-1}, y_T.$$

5.1.2 Distributions

We will think of $\mathbf{Y}_T = \{Y_t\}_{t=1}^T$ as a random variable in its own right.

- $\mathbf{y}_T = \{y_t\}_{t=1}^T$ is a *single* realization of $\mathbf{Y}_T = \{Y_t\}_{t=1}^T$.
- The CDF is $F_{\mathbf{Y}_T}(\mathbf{y}_T)$ and the PDF is $f_{\mathbf{Y}_T}(\mathbf{y}_T)$.
- For example, consider $T = 100$:

$$F(\mathbf{y}_{100}) = P(Y_1 \leq y_1, \ldots, Y_{100} \leq y_{100}).$$

- Notice that $\mathbf{Y}_T$ is just a collection of random variables and $f_{\mathbf{Y}_T}(\mathbf{y}_T)$ is the joint density.

5.1.3 Time Series Observations

As statisticians and econometricians, we want many observations of $\mathbf{Y}_T$ to learn about its distribution:

$$\mathbf{y}_T^{(1)}, \quad \mathbf{y}_T^{(2)}, \quad \mathbf{y}_T^{(3)}, \quad \ldots$$

Likewise, if we are only interested in the marginal distribution of Y_{17}

$$f_{Y_{17}}(a) = P(Y_{17} \leq a)$$

we want many observations: $\left\{y_{17}^{(i)}\right\}_{i=1}^N$.

5.1.4 Time Series Observations

Unfortunately, we usually only have *one observation* of $\mathbf{Y}_T$.

- Think of the daily closing price of Harley-Davidson stock since January 2nd.
- Think of your cardiogram for the past 100 seconds.

In neither case can you repeat history to observe a new sequence of prices or electronic heart signals.

- In time series econometrics we typically base inference on a single observation.
- Additional assumptions about the process will allow us to exploit information in the full sequence $\mathbf{y}_T$ to make inferences about the joint distribution $F_{\mathbf{Y}_T}(\mathbf{y}_T)$.

5.1.5 Moments

Since the stochastic process is comprised of individual random variables, we can consider moments of each:

$$E[Y_t] = \int_{-\infty}^{\infty} y_t f_{Y_t}(y_t) dy_t = \mu_t$$

$$Var(Y_t) = \int_{-\infty}^{\infty} (y_t - \mu_t)^2 f_{Y_t}(y_t) dy_t = \gamma_{0t}$$

$$Cov(Y_t, Y_{t-j}) = \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} (y_t - \mu_t)(y_{t-j} - \mu_{t-j}) \times f_{Y_t, Y_{t-j}}(y_t, y_{t-j}) dy_t dy_{t-j} = \gamma_{jt},$$

where f_{Y_t} and $f_{Y_t, Y_{t-j}}$ are the marginal distributions of $f_{\mathbf{Y}_T}$ obtained by integrating over the appropriate elements of $\mathbf{Y}_T$.

5.1.6 Autocovariance and Autocorrelation

- γ_{jt} is known as the j th autocovariance of Y_t since it is the covariance of Y_t with its own lagged value.
- The j th autocorrelation of Y_t is defined as

$$\rho_{jt} = Corr(Y_t, Y_{t-j})$$

$$= \frac{Cov(Y_t, Y_{t-j})}{\sqrt{Var(Y_t)}\sqrt{Var(Y_{t-j})}}$$

$$= \frac{\gamma_{jt}}{\sqrt{\gamma_{0t}}\sqrt{\gamma_{0t-j}}}.$$

5.1.7 Sample Moments

If we had N observations $\mathbf{y}_T^{(1)}, \ldots, \mathbf{y}_T^{(N)}$, we could estimate moments of each (univariate) Y_t in the usual way:

$$\hat{\mu}_t = \frac{1}{N} \sum_{i=1}^{N} y_t^{(i)}.$$

$$\hat{\gamma}_{0t} = \frac{1}{N}\sum_{i=1}^{N}(y_t^{(i)} - \hat{\mu}_t)^2.$$

$$\hat{\gamma}_{jt} = \frac{1}{N}\sum_{i=1}^{N}(y_t^{(i)} - \hat{\mu}_t)(y_{t-j}^{(i)} - \hat{\mu}_{t-j}).$$

5.1.8 Example

Suppose each element of $\mathbf{Y}_T$ is described by

$$Y_t = \mu_t + \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, \sigma_t^2), \forall t.$$

5.1.9 Example

In this case,

$$\mu_t = E[Y_t] = \mu_t, \ \ \forall t,$$

$$\gamma_{0t} = Var(Y_t) = Var(\varepsilon_t) = \sigma_t^2, \ \ \forall t$$

$$\gamma_{jt} = Cov(Y_t, Y_{t-j}) = Cov(\varepsilon_t, \varepsilon_{t-j}) = 0, \ \ \forall t, j \neq 0.$$

- If $\sigma_t^2 = \sigma^2 \ \forall t$, ε_T is known as a *Gaussian white noise* process.
- In this case, $\mathbf{Y}_T$ is a Gaussian white noise process with drift.
- μ_T is the drift vector.

5.1.10 White Noise

Generally speaking, ε_T is a *white noise process* if

$$E[\varepsilon_t] = 0, \ \ \forall t \tag{5.1}$$

$$E[\varepsilon_t^2] = \sigma^2, \ \ \forall t \tag{5.2}$$

$$E[\varepsilon_t \varepsilon_\tau] = 0, \quad \text{for } t \neq \tau. \tag{5.3}$$

5.1.11 White Noise

Notice there is no distributional assumption for ε_t.

- If ε_t and ε_τ are independent for $t \neq \tau$, ε_T is *independent white noise*.
- Notice that independence $\Rightarrow$ Equation (5.3), but Equation (5.3) does not $\Rightarrow$ independence.
- If $\varepsilon_t \sim \mathcal{N}(0, \sigma^2) \ \forall t$, as in the example above, ε_T is Gaussian white noise.

5.1.12 Weak Stationarity

Suppose the first and second moments of a stochastic process $\mathbf{Y}_T$ don't depend on $t \in T$:

$$E[Y_t] = \mu \quad \forall t$$

$$Cov(Y_t, Y_{t-j}) = \gamma_j \quad \forall t \text{ and any } j.$$

- In this case $\mathbf{Y}_T$ is *weakly stationary* or *covariance stationary*.
- In the previous example, if $Y_t = \mu + \varepsilon_t \; \forall t$, $\mathbf{Y}_T$ is weakly stationary.
- However if $\mu_t \neq \mu \; \forall t$, $\mathbf{Y}_T$ is *not* weakly stationary.

5.1.13 Autocorrelation under Weak Stationarity

If $\mathbf{Y}_T$ is weakly stationary

$$\rho_{jt} = \frac{\gamma_{jt}}{\sqrt{\gamma_{0t}}\sqrt{\gamma_{0t-j}}}$$

$$= \frac{\gamma_j}{\sqrt{\gamma_0}\sqrt{\gamma_0}}$$

$$= \frac{\gamma_j}{\gamma_0}$$

$$= \rho_j.$$

- Note that $\rho_0 = 1$.

5.1.14 Weak Stationarity

Under weak stationarity, autocovariances γ_j only depend on the distance between random variables within a stochastic process:

$$Cov(Y_\tau, Y_{\tau-j}) = Cov(Y_t, Y_{t-j}) = \gamma_j.$$

This implies

$$\gamma_{-j} = Cov(Y_{t+j}, Y_t) = Cov(Y_t, Y_{t-j}) = \gamma_j.$$

5.1.15 Weak Stationarity

More generally,

$$\Sigma_{\mathbf{Y}_T} = \begin{bmatrix} \gamma_0 & \gamma_1 & \cdots & \gamma_{T-2} & \gamma_{T-1} \\ \gamma_1 & \gamma_0 & \cdots & \gamma_{T-3} & \gamma_{T-2} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \gamma_{T-2} & \gamma_{T-3} & \cdots & \gamma_0 & \gamma_1 \\ \gamma_{T-1} & \gamma_{T-2} & \cdots & \gamma_1 & \gamma_0 \end{bmatrix}.$$

5.1.16 Strict Stationarity

$\mathbf{Y}_T$ is *strictly stationary* if for any set $\{j_1, j_2, \ldots, j_n\} \in T$

$$f_{Y_{j_1},\ldots,Y_{j_n n}}(a_1, \ldots, a_n) = f_{Y_{j_1+\tau},\ldots,Y_{j_n n+\tau}}(a_1, \ldots, a_n), \quad \forall \tau.$$

- Strict stationarity means that the joint distribution of any subset of random variables in $\mathbf{Y}_T$ is invariant to shifts in time, τ.
- Strict stationarity $\Rightarrow$ weak stationarity if the first and second moments of a stochastic process exist.
- Weak stationarity does note $\Rightarrow$ strict stationarity: invariance of first and second moments to time shifts (weak stationarity) does not mean that all higher moments are invariant to time shifts (strict stationarity).

5.1.17 Strict Stationarity

If $\mathbf{Y}_T$ is Gaussian then weak stationarity $\Rightarrow$ strict stationarity.

- If $\mathbf{Y}_T$ is Gaussian, all marginal distributions of $(Y_{j_1}, \ldots, Y_{j_n})$ are also Gaussian.
- Gaussian distributions are fully characterized by their first and second moments.

5.2 Autoregressive Processes

5.2.1 $AR(1)$ Process

Given white noise $\{\varepsilon_t\}$, consider the process

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t,$$

where c and ϕ are constants.

- This is a *first-order autoregressive* or $AR(1)$ process.
- ϕ can be thought of as a *memory* or *feedback* parameter and introduces serial correlation in Y_t.
- When $\phi = 0$, Y_t is white noise with drift - it has no memory or serial correlation.

5.2.2 Recursive Substitution of $AR(1)$

Applying recursive substitution:

$$\begin{aligned} Y_t &= c + \phi Y_{t-1} + \varepsilon_t \\ &= c + \phi(c + \phi Y_{t-2} + \varepsilon_{t-1}) + \varepsilon_t \\ &= c + \phi c + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2 Y_{t-2} \\ &= c + \phi c + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2(c + \phi Y_{t-3} + \varepsilon_{t-2}) \end{aligned}$$

5.2.3 Recursive Substitution of $AR(1)$

$$= c + \phi c + \phi^2 c + \varepsilon_t + \phi\varepsilon_{t-1} + \phi^2\varepsilon_{t-2} + \phi^3 Y_{t-3}$$

$$\vdots$$

$$= \sum_{i=0}^{\infty} \phi^i c + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}.$$

$$= \frac{c}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}.$$

- The infinite recursive substitution can only be performed if $|\phi| < 1$.

5.2.4 Expectation of $AR(1)$

Assume Y_t is weakly stationary: $|\phi| < 1$.

$$E[Y_t] = c + \phi E[Y_{t-1}] + E[\varepsilon_t]$$

$$= c + \phi E[Y_t]$$

$$\Rightarrow E[Y_t] = \frac{c}{1-\phi}.$$

5.2.5 A Useful Property

If Y_t is weakly stationary,

$$Y_{t-j} - \mu = \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-j-i}.$$

- That is, for $j \geq 1$, Y_{t-j} is a function of lagged values of ε_t and not ε_t itself.
- As a result, for $j \geq 1$

$$E\left[(Y_{t-j} - \mu)\varepsilon_t\right] = \sum_{i=0}^{\infty} \phi^i E[\varepsilon_t \varepsilon_{t-j-i}] = 0.$$

5.2.6 Variance of $AR(1)$

Given that $\mu = c/(1-\phi)$ for weakly stationary Y_t:

$$Y_t = \mu(1-\phi) + \phi Y_{t-1} + \varepsilon_t$$

$$\Rightarrow (Y_t - \mu) = \phi(Y_{t-1} - \mu) + \varepsilon_t.$$

Squaring both sides and taking expectations:

$$E\left[(Y_t-\mu)^2\right] = \phi^2 E\left[(Y_{t-1}-\mu)^2\right] + 2\phi E\left[(Y_{t-1}-\mu)\varepsilon_t\right] + E[\varepsilon_t^2]$$

$$= \phi^2 E\left[(Y_t-\mu)^2\right] + \sigma^2$$

$$\Rightarrow (1-\phi^2)\gamma_0 = \sigma^2$$

$$\Rightarrow \gamma_0 = \frac{\sigma^2}{1-\phi^2}$$

5.2.7 Autocovariances of $AR(1)$

For $j \geq 1$,

$$\gamma_j = E\left[(Y_t-\mu)(Y_{t-j}-\mu)\right]$$

$$= \phi E[(Y_{t-1}-\mu)(Y_{t-j}-\mu)] + E[\varepsilon_t(Y_{t-j}-\mu)]$$

$$= \phi\gamma_{j-1}$$

$$\vdots$$

$$= \phi^j\gamma_0.$$

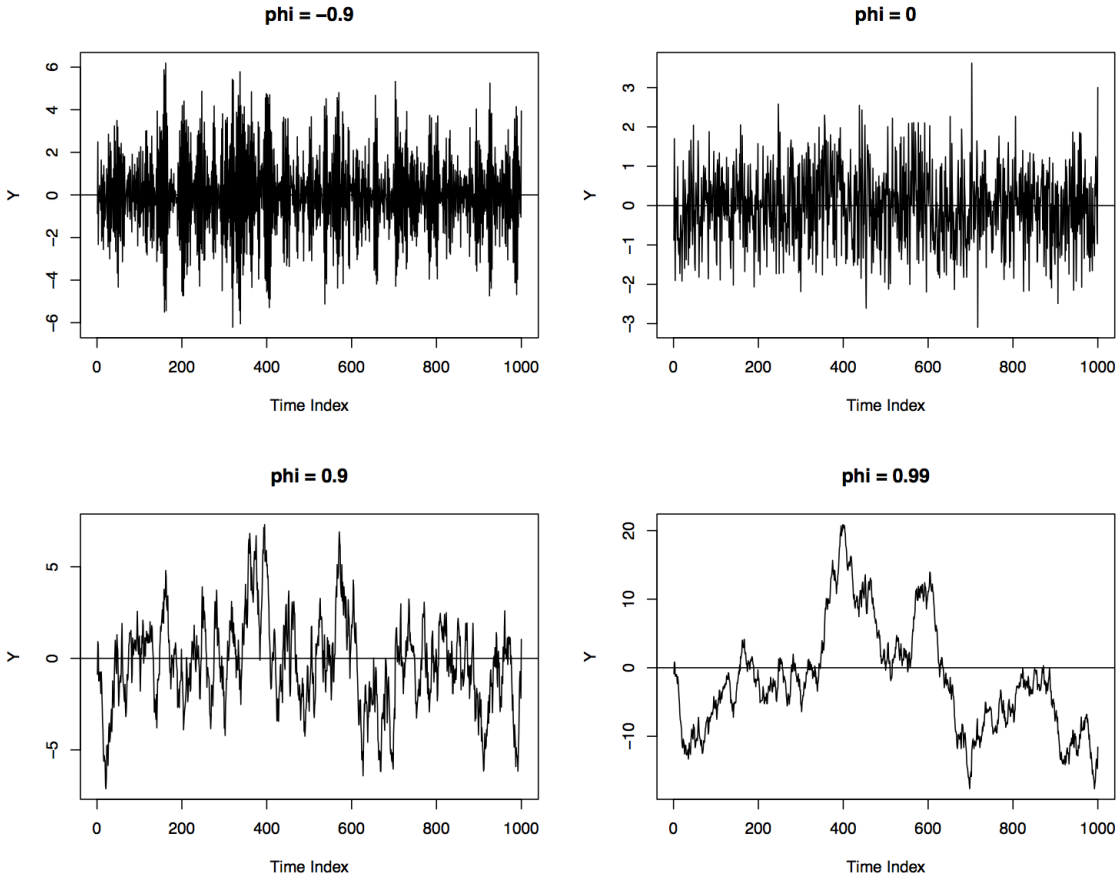
5.2.8 Autocorrelations of $AR(1)$

The autocorrelations of an $AR(1)$ are

$$\rho_j = \frac{\gamma_j}{\gamma_0} = \phi^j, \quad \forall j \geq 0.$$

- Since we assumed $|\phi| < 1$, the autocorrelations decay exponentially as j increases.
- Note that if $\phi \in (-1, 0)$, the autocorrelations decay in an oscillatory fashion.

5.2.9 Examples of $AR(1)$ Processes

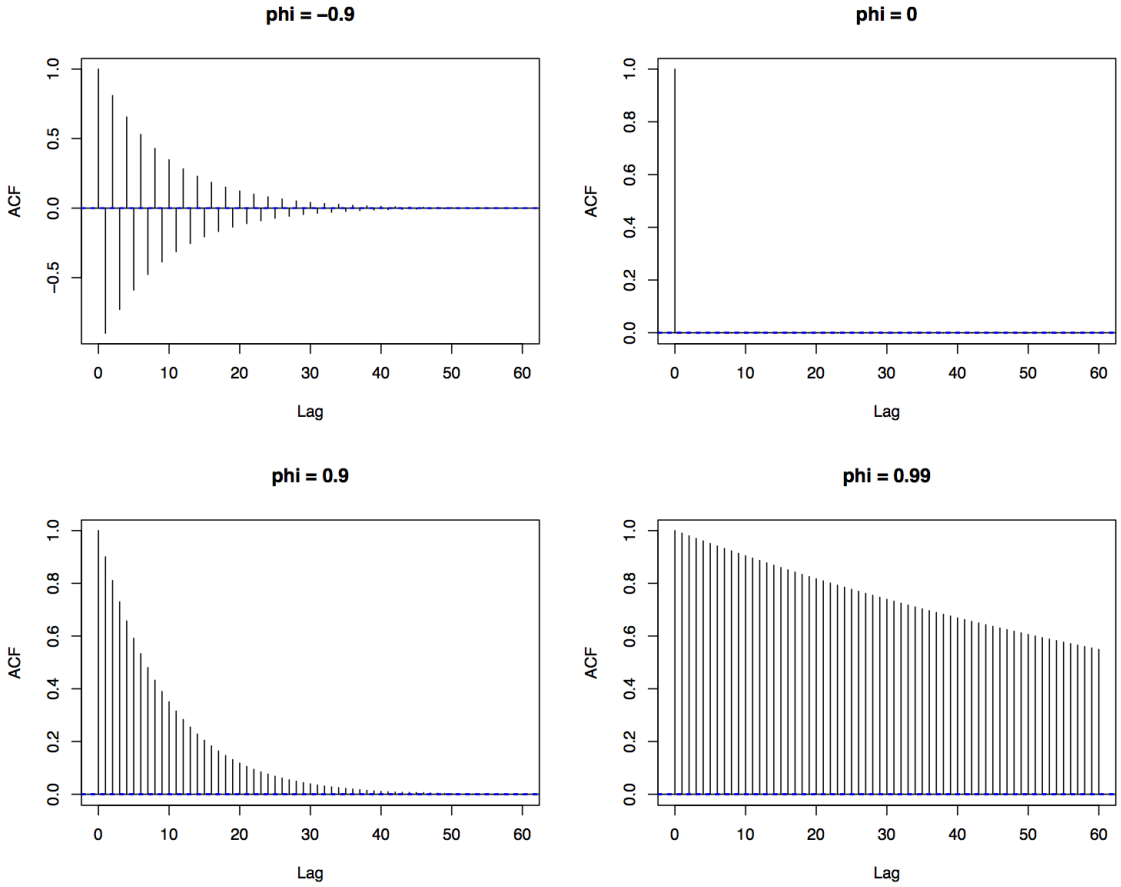


To create this plot, run the following script:

```
############################################################
# Simulate AR(1) processes for different values of phi
############################################################

# Number of simulated points
nSim = 1000000;

# Values of phi to consider
#phi = c(-0.9, 0, 0.9, 0.99);
phi = c(0.98, 0.6, 1, 1.01);

# Draw one set of shocks and use for each AR(1)
eps = rnorm(nSim, 0, 1);

# Matrix which stores each AR(1) in columns
y = matrix(0, nrow=nSim, ncol=length(phi));

# Each process is intialized at first shock
y[1,] = eps[1];

# Loop over each value of phi
for(j in 1:length(phi)){
```

```
  # Loop through the series, simulating the AR(1) values
  for(i in 2:nSim){
    y[i,j] = phi[j]*y[i-1,j]+eps[i]
  }
}

############################################################
# Plot the AR(1) realizations for each phi
############################################################

# Only plot a subset of the whole simulation
plotInd = 1:200

# Specify a plot grid
pdf(file="ar1ExampleSeriesAlt.pdf", height=8, width=10)
par(mfrow=c(2,2))

# Loop over each value of phi
for(j in 1:length(phi)){
  plot(plotInd,y[plotInd,j], type='l', xlab='Time Index',
       ylab="Y", main=paste(expression(phi), " = ", phi[j], sep=""))
  abline(h=0)
}
graphics.off()
```

5.2.10 $AR(1)$ Autocorrelations



To create this plot, run the following script:

```
############################################################
# Plot the sample ACFs for each AR(1) simulation
# For large nSim, sample ACFs are close to true ACFs
############################################################

# Specify a plot grid
pdf(file="ar1ExampleACFAlt.pdf", height=8, width=10)
par(mfrow=c(2,2))

# Loop over each value of phi
for(j in 1:length(phi)){
  acf(y[is.finite(y[,j]),j], main=paste(expression(phi), " = ", phi[j], sep=""))
}
graphics.off()
```

5.2.11 Random Walk

Suppose $\phi = 1$:

$$Y_t = c + Y_{t-1} + \varepsilon_t = \cdots = tc + Y_0 + \varepsilon_1 + \cdots \varepsilon_t.$$

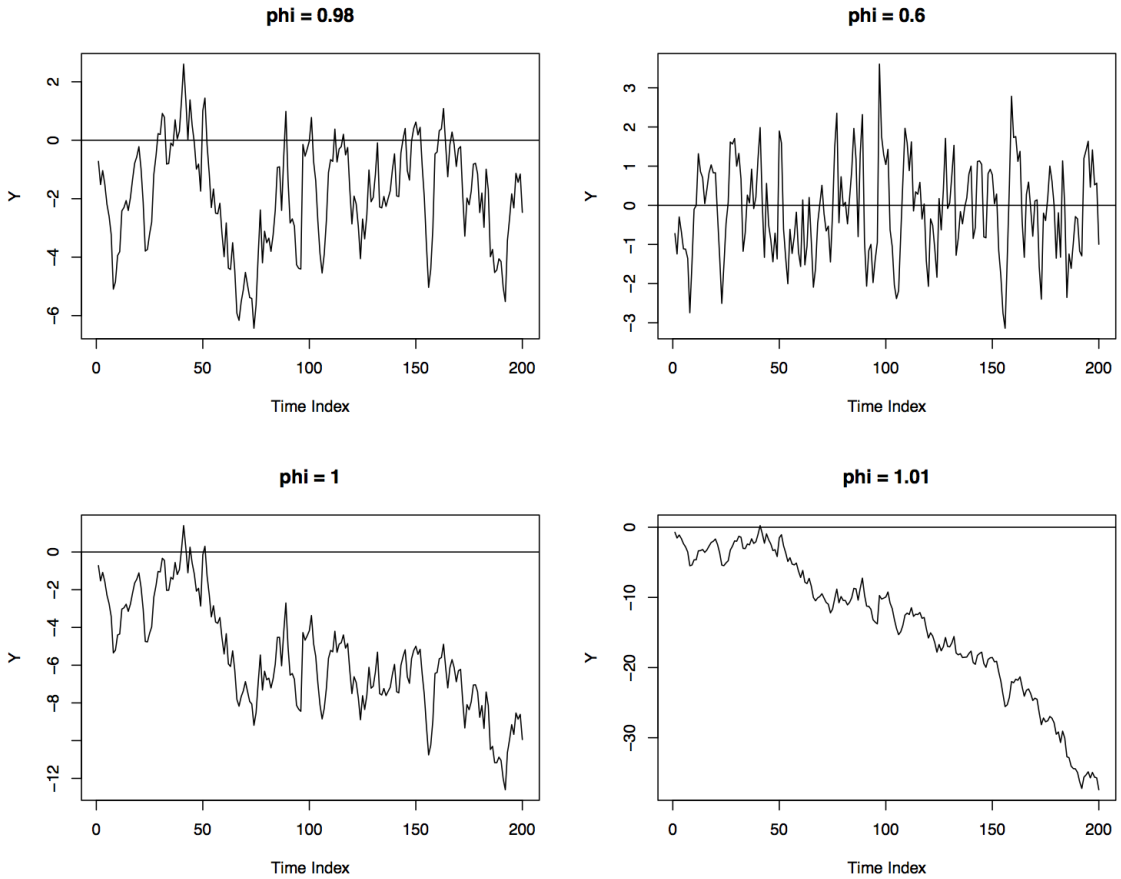
- Clearly $E[Y_t] = tc + Y_0$, which is not independent of time.
- $Var(Y_t) = t\sigma^2$, which increases linearly with time.

5.2.12 Explosive $AR(1)$

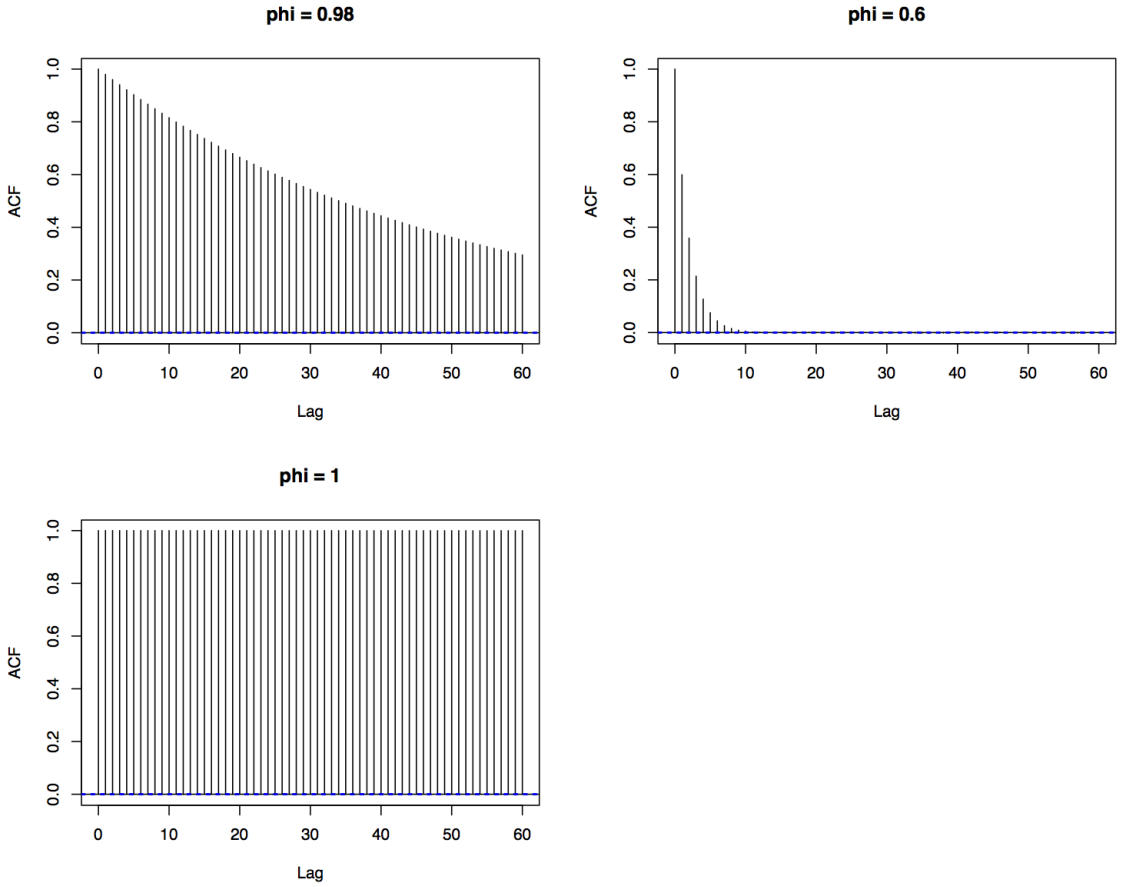
When $|\phi| > 1$, the autoregressive process is explosive.

- Recall that $Y_t = \frac{c}{1-\phi} + \sum_{i=0}^{\infty} \phi^i \varepsilon_{t-i}$.
- Now $|\phi^i|$ increases with i rather than decay.
- Past values of ε_{t-i} contribute greater amounts to Y_t as i increases.

5.2.13 Examples of $AR(1)$ Processes



5.2.14 $AR(1)$ Autocorrelations



5.2.15 $AR(p)$ Process

Given white noise $\{\varepsilon_t\}$, consider the process

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \ldots + \phi_p Y_{t-p} + \varepsilon_t,$$

where c and $\{\phi\}_{i=1}^{p}$ are constants.

- This is a *p th-order autoregressive* or $AR(p)$ process.

5.2.16 Expectation of $AR(p)$

Assume Y_t is weakly stationary.

$$E[Y_t] = c + \phi_1 E[Y_{t-1}] + \ldots + \phi_p E[Y_{t-p}] + E[\varepsilon_t]$$

$$= c + \phi_1 E[Y_t] + \ldots + \phi_p E[Y_t]$$

$$\Rightarrow E[Y_t] = \frac{c}{1 - \phi_1 - \ldots - \phi_p} = \mu.$$

5.2.17 Autocovariances of $AR(p)$

Given that $\mu = c/(1 - \phi_1 - \ldots - \phi_p)$ for weakly stationary Y_t:

$$Y_t = \mu(1 - \phi_1 - \ldots - \phi_p) + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t$$

$$\Rightarrow (Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \ldots + \phi_p(Y_{t-p} - \mu) + \varepsilon_t.$$

5.2.18 Autocovariances of $AR(p)$

Thus,

$$\gamma_j = E\left[(Y_t - \mu)(Y_{t-j} - \mu)\right]$$

$$= \phi_1 E\left[(Y_{t-1} - \mu)(Y_{t-j} - \mu)\right] + \ldots$$

$$+ \phi_p E\left[(Y_{t-p} - \mu)(Y_{t-j} - \mu)\right] + E\left[\varepsilon_t(Y_{t-j} - \mu)\right]$$

$$= \begin{cases} \phi_1\gamma_{j-1} + \ldots + \phi_p\gamma_{j-p} & \text{for } j = 1, \ldots \\ \phi_1\gamma_1 + \ldots + \phi_p\gamma_p + \sigma^2 & \text{for } j = 0. \end{cases} \tag{5.4}$$

5.2.19 Autocovariances of $AR(p)$

For $j = 0, 1, \ldots, p$, System (5.4) is a system of $p + 1$ equations with $p + 1$ unknowns: $\{\gamma_j\}_{j=0}^p$.

- $\{\gamma_j\}_{j=0}^p$ can be solved for as functions of $\{\phi_j\}_{j=1}^p$ and σ^2.
- It can be shown that $\{\gamma_j\}_{j=0}^p$ are the first p elements of the first column of $\sigma^2[I_{p^2} - \Phi \otimes \Phi]^{-1}$, where $\otimes$ denotes the Kronecker product.
- $\{\gamma_j\}_{j=p+1}^\infty$ can then be determined using prior values of γ_j and $\{\phi_j\}_{j=1}^p$.

5.2.20 Autocorrelations of $AR(p)$

Dividing the autocovariances by γ_0,

$$\rho_j = \phi_1\rho_{j-1} + \ldots + \phi_p\rho_{j-p} \quad \text{for } j = 1, 2, 3, \ldots$$

5.2.21 Estimating AR Models

Ideally, estimation of an AR model is done via maximum likelihood.

- For an $AR(p)$ model, one would first specify a joint likelihood for the parameters $\phi_1, \ldots, \phi_p, c, \sigma^2$.
- Taking derivatives of the log likelihood with respect to each of the parameters would result in a system of equations that could be solved for the MLEs: $\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{c}, \hat{\sigma}^2$.

5.2.22 Estimating AR Models

- The exact likelihood is a bit cumbersome and maximization requires specialized numerical methods.
- It turns out that the least squares estimates obtained by fitting a regression of Y_t on $Y_{t-1}, \ldots, Y_{t-p}$ are almost identical to the MLEs (they are called the conditional MLEs).

5.2.23 Estimating AR Models

- The exact MLEs can be obtained with the `arima` function in `R`.
- The conditional (least squares) MLEs can be obtained with the `lm` function in `R`.

5.2.24 Which AR?

How do we know if an AR model is appropriate and which AR model to fit?

- After fitting an AR model, we can examine the residuals.
- The `acf` function in `R` can be used to compute empirical autocorrelations of the residuals.
- If the residuals are autocorrelated, the model is not a good fit. Consider increasing the order of the AR or using another model.

5.2.25 Which AR?

Suppose Y_t is an $AR(2)$ process:

$$(Y_t - \mu) = \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \varepsilon_t.$$

- If we estimate an $AR(1)$ model using the data, then for large sample sizes $\hat{\mu} \approx \mu$ and $\hat{\phi} \approx E[\hat{\phi}] = \phi^* \neq \phi_1$.

5.2.26 Which AR?

The resulting residuals would be

$$\begin{aligned}\hat{\varepsilon}_t &= (Y_t - \mu) - \phi^*(Y_{t-1} - \mu) \\ &= \phi_1(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \varepsilon_t - \phi^*(Y_{t-1} - \mu) \\ &= (\phi_1 - \phi^*)(Y_{t-1} - \mu) + \phi_2(Y_{t-2} - \mu) + \varepsilon_t - \phi^*(Y_{t-1} - \mu).\end{aligned}$$

- Even if $\phi^* = \phi_1$, the residuals will exhibit autocorrelation, due to the presence of Y_{t-2}.

5.2.27 Which AR?

The `auto.arima` function in `R` estimates a range of $AR(p)$ models and selects the one with the best fit.

- `auto.arima` uses the *Akaike Information Criterion* (AIC) or the *Bayesian Information Criterion* (BIC) to select the model.
- Minimizing AIC and BIC amounts to minimizing the sum of squared residuals, with a penalty term that is related to the number of model parameters.

5.3 Moving Average Processes

5.3.1 $MA(1)$

Given white noise $\{\varepsilon_t\}$, consider the process

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1},$$

where μ and θ are constants.

- This is a *first-order moving average* or $MA(1)$ process.

5.3.2 $MA(1)$ Mean and Variance

The mean of the first-order moving average process is

$$\begin{aligned} E[Y_t] &= E[\mu + \varepsilon_t + \theta\varepsilon_{t-1}] \\ &= \mu + E[\varepsilon_t] + \theta E[\varepsilon_{t-1}] \\ &= \mu. \end{aligned}$$

5.3.3 $MA(1)$ Autocovariances

$$\begin{aligned} \gamma_j &= E\left[(Y_t - \mu)(Y_{t-j} - \mu)\right] \\ &= E\left[(\varepsilon_t + \theta\varepsilon_{t-1})(\varepsilon_{t-j} + \theta\varepsilon_{t-j-1})\right] \\ &= E[\varepsilon_t\varepsilon_{t-j} + \theta\varepsilon_t\varepsilon_{t-j-1} + \theta\varepsilon_{t-1}\varepsilon_{t-j} + \theta^2\varepsilon_{t-1}\varepsilon_{t-j-1}] \\ &= E[\varepsilon_t\varepsilon_{t-j}] + \theta E[\varepsilon_t\varepsilon_{t-j-1}] + \theta E[\varepsilon_{t-1}\varepsilon_{t-j}] + \theta^2 E[\varepsilon_{t-1}\varepsilon_{t-j-1}]. \end{aligned}$$

5.3.4 $MA(1)$ Autocovariances

- If $j = 0$

$$\gamma_0 = E[\varepsilon_t^2] + \theta E[\varepsilon_t\varepsilon_{t-1}] + \theta E[\varepsilon_{t-1}\varepsilon_t] + \theta^2 E[\varepsilon_{t-1}^2] = (1+\theta^2)\sigma^2.$$

- If $j = 1$

$$\gamma_1 = E[\varepsilon_t\varepsilon_{t-1}] + \theta E[\varepsilon_t\varepsilon_{t-2}] + \theta E[\varepsilon_{t-1}^2] + \theta^2 E[\varepsilon_{t-1}\varepsilon_{t-2}] = \theta\sigma^2.$$

- If $j > 1$, all of the expectations are zero:

$$\gamma_j = 0.$$

5.3.5 $MA(1)$ Stationarity

Since the mean and autocovariances are independent of time, an $MA(1)$ is weakly stationary.

- This is true *for all values of* θ.

5.3.6 $MA(1)$ Autocorrelations

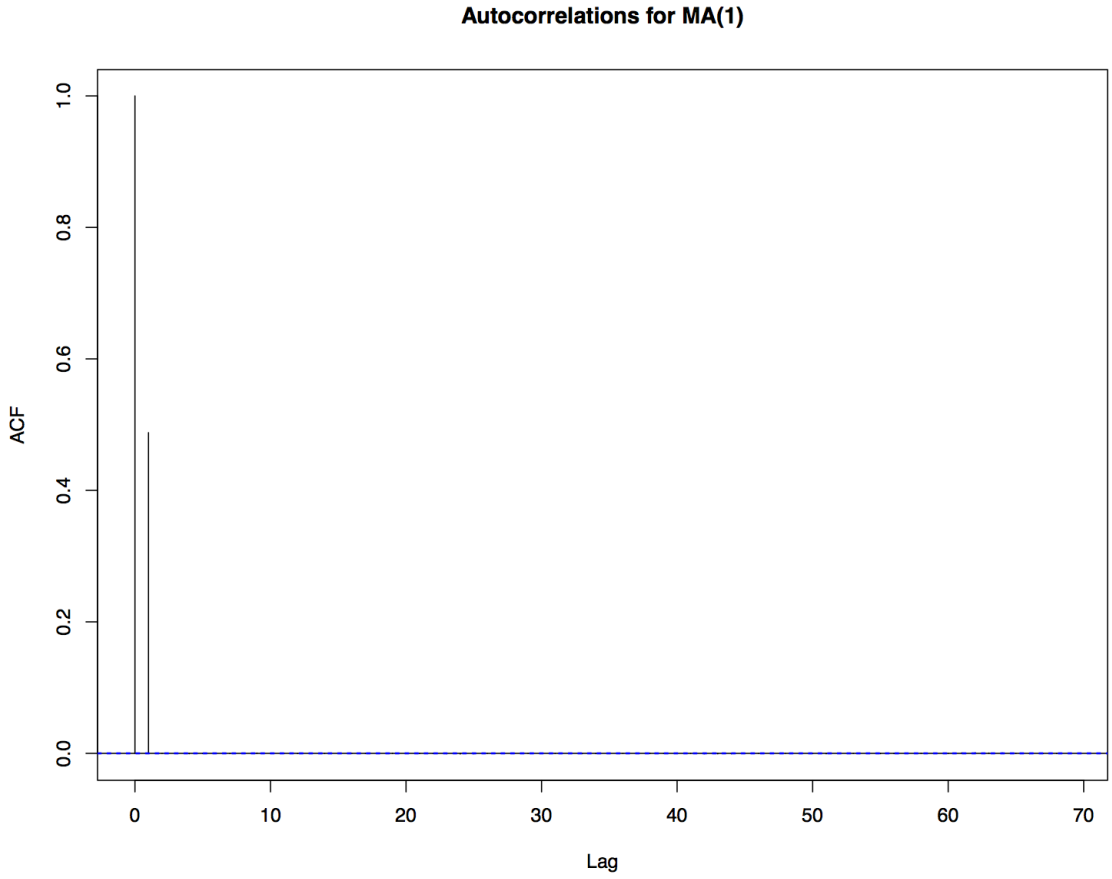
The autocorrelations of an $MA(1)$ are

- $j = 0$: $\rho_0 = 1$ (*always*).
- $j = 1$:

$$\rho_1 = \frac{\theta\sigma^2}{(1+\theta^2)\sigma^2} = \frac{\theta}{1+\theta^2}$$

- $j > 1$: $\rho_j = 0$.
- If $\theta > 0$, first-order lags of Y_t are *positively* autocorrelated.
- If $\theta < 0$, first-order lags of Y_t are *negatively* autocorrelated.

5.3.7 $MA(1)$ Autocorrelations



5.3.8 $MA(q)$

A *q th-order moving average* or $MA(q)$ process is

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q},$$

where $\mu, \theta_1, \ldots, \theta_q$ are any real numbers.

5.3.9 $MA(q)$ Mean

As with the $MA(1)$:

$$\begin{aligned} E[Y_t] &= E[\mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}] \\ &= \mu + E[\varepsilon_t] + \theta_1 E[\varepsilon_{t-1}] + \ldots + \theta_q E[\varepsilon_{t-q}] \\ &= \mu. \end{aligned}$$

5.3.10 $MA(q)$ Autocovariances

$$\begin{aligned} \gamma_j &= E\left[(Y_t - \mu)(Y_{t-j} - \mu)\right] \\ &= E\big[(\varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots + \theta_q \varepsilon_{t-q}) \\ &\qquad \times (\varepsilon_{t-j} + \theta_1 \varepsilon_{t-j-1} + \ldots + \theta_q \varepsilon_{t-j-q})\big]. \end{aligned}$$

- For $j > q$, all of the products result in zero expectations: $\gamma_j = 0$, for $j > q$.

5.3.11 $MA(q)$ Autocovariances

- For $j = 0$, the squared terms result in nonzero expectations, while the cross products lead to zero expectations:

$$\gamma_0 = E[\varepsilon_t^2] + \theta_1^2 E[\varepsilon_{t-1}^2] + \ldots + \theta_q^2 E[\varepsilon_{t-q}^2] = \left(1 + \sum_{j=1}^{q} \theta_j^2\right) \sigma^2.$$

5.3.12 $MA(q)$ Autocovariances

- For $j = \{1, 2, \ldots, q\}$, the nonzero expectation terms are

$$\begin{aligned} \gamma_j &= \theta_j E[\varepsilon_{t-j}^2] + \theta_{j+1}\theta_1 E[\varepsilon_{t-j-1}^2] \\ &\qquad + \theta_{j+2}\theta_2 E[\varepsilon_{t-j-2}^2] + \ldots + \theta_q \theta_{q-j} E[\varepsilon_{t-q}^2] \\ &= (\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \ldots + \theta_q \theta_{q-j})\sigma^2. \end{aligned}$$

The autocovariances can be stated concisely as

$$\gamma_j = \begin{cases} (\theta_j + \theta_{j+1}\theta_1 + \theta_{j+2}\theta_2 + \ldots + \theta_q \theta_{q-j})\sigma^2 & \text{for } j = 0, 1, \ldots, q \\ 0 & \text{for } j > q. \end{cases}$$

where $\theta_0 = 1$.

5.3.13 $MA(q)$ Autocorrelations

The autocorrelations can be stated concisely as

$$\rho_j = \begin{cases} \frac{\theta_j+\theta_{j+1}\theta_1+\theta_{j+2}\theta_2+\ldots+\theta_q\theta_{q-j}}{\theta_0^2+\theta_1^2+\theta_2^2+\ldots+\theta_q^2} & \text{for } j = 0, 1, \ldots, q \\ 0 & \text{for } j > q. \end{cases}$$

where $\theta_0 = 1$.

5.3.14 $MA(2)$ Example

For an $MA(2)$ process

$$\gamma_0 = (1 + \theta_1^2 + \theta_2^2)\sigma^2$$

$$\gamma_1 = (\theta_1 + \theta_2\theta_1)\sigma^2$$

$$\gamma_2 = \theta_2\sigma^2$$

$$\gamma_3 = \gamma_4 = \ldots = 0.$$

5.3.15 Estimating MA Models

Estimation of an MA model is done via maximum likelihood.

- For an $MA(q)$ model, one would first specify a joint likelihood for the parameters $\{\theta_1, \ldots, \theta_q, \mu, \sigma^2\}$.
- Taking derivatives of the log likelihood with respect to each of the parameters would result in a system of equations that could be solved for the MLEs: $\{\hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\mu}, \hat{\sigma}^2\}$.
- The exact likelihood is a bit cumbersome and maximization requires specialized numerical methods.
- The MLEs can be obtained with the `arima` function in `R`.

5.3.16 Which MA?

How do we know if an MA model is appropriate and which MA model to fit?

- For an $MA(q)$, we know that $\gamma_j = 0$ for $j > q$.
- We should only fit an MA model if the autocorrelations drop to zero for all $j > q$ for some value q.
- The `acf` function in `R` can be used to compute empirical autocorrelations of the data.
- The appropriate q can then be obtained from the empirical ACF.

5.3.17 Which MA?

- After fitting an MA model, we can examine the residuals.
- The `acf` function can be used to compute empirical autocorrelations of the residuals.
- If the residuals are autocorrelated, the model is not a good fit. Consider changing the order of the MA or using another model.

5.3.18 Which MA?

The `auto.arima` function in `R` estimates a range of $MA(q)$ models and selects the one with the best fit.

- `auto.arima` uses the *Akaike Information Criterion* (AIC) or the *Bayesian Information Criterion* (BIC) to select the model.
- Minimizing AIC and BIC amounts to minimizing the sum of squared residuals, with a penalty term that is related to the number of model parameters.

5.4 ARMA Processes

5.4.1 $ARMA(p,q)$ Process

Given white noise $\{\varepsilon_t\}$, consider the process

$$Y_t = c + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \ldots \theta_q \varepsilon_{t-q},$$

where c, $\{\phi_i\}_{i=1}^p$ and $\{\theta_i\}_{i=1}^q$ are constants.

- This is an $ARMA(p,q)$ process.

5.4.2 Expectation of $ARMA(p,q)$

Assume Y_t is weakly stationary.

$$\begin{aligned} E[Y_t] &= c + \phi_1 E[Y_{t-1}] + \ldots + \phi_p E[Y_{t-p}] \\ &\qquad + E[\varepsilon_t] + \theta_1 E[\varepsilon_{t-1}] + \ldots + \theta_q E[\varepsilon_{t-q}] \\ &= c + \phi_1 E[Y_t] + \ldots + \phi_p E[Y_t] \\ \Rightarrow E[Y_t] &= \frac{c}{1 - \phi_1 - \ldots - \phi_p} = \mu. \end{aligned}$$

- This is the same mean as an $AR(p)$ process with parameters c and $\{\phi_i\}_{i=1}^p$.

5.4.3 Autocovariances of $ARMA(p,q)$

Given that $\mu = c/(1 - \phi_1 - \ldots - \phi_p)$ for weakly stationary Y_t:

$$\begin{aligned} Y_t &= \mu(1 - \phi_1 - \ldots - \phi_p) + \phi_1 Y_{t-1} + \ldots + \phi_p Y_{t-p} \\ &\qquad + \varepsilon_t + \theta_1 \varepsilon_1 + \ldots \theta_q \varepsilon_{t-q} \\ \Rightarrow (Y_t - \mu) &= \phi_1 (Y_{t-1} - \mu) + \ldots + \phi_p (Y_{t-p} - \mu) \\ &\qquad + \varepsilon_t + \theta_1 \varepsilon_1 + \ldots \theta_q \varepsilon_{t-q}. \end{aligned}$$

$$\begin{aligned} \gamma_j &= E\left[(Y_t - \mu)(Y_{t-j} - \mu)\right] \\ &= \phi_1 E\left[(Y_{t-1} - \mu)(Y_{t-j} - \mu)\right] + \ldots \\ &\qquad + \phi_p E\left[(Y_{t-p} - \mu)(Y_{t-j} - \mu)\right] \\ &\qquad + E\left[\varepsilon_t (Y_{t-j} - \mu)\right] + \theta_1 E\left[\varepsilon_{t-1}(Y_{t-j} - \mu)\right] \\ &\qquad + \ldots + \theta_q E\left[\varepsilon_{t-q}(Y_{t-j} - \mu)\right] \end{aligned}$$

5.4.4 Autocovariances of $ARMA(p, q)$

- For $j > q$, γ_j will follow the same law of motion as for an $AR(p)$ process:

$$\gamma_j = \phi_1\gamma_{j-1} + \ldots + \phi_p\gamma_{j-p} \quad \text{for } j = q+1, \ldots$$

- These values will not be the same as the $AR(p)$ values for $j = q+1, \ldots$, since the initial $\gamma_0, \ldots, \gamma_q$ will differ.
- The first q autocovariances, $\gamma_0, \ldots, \gamma_q$, of an $ARMA(p, q)$ will be more complicated than those of an $AR(p)$.

5.4.5 Estimating $ARMA$ Models

Estimation of an $ARMA$ model is done via maximum likelihood.

- For an $ARMA(p, q)$ model, one would first specify a joint likelihood for the parameters $\{\phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \mu, \sigma^2\}$.
- Taking derivatives of the log likelihood with respect to each of the parameters would result in a system of equations that could be solved for the MLEs: $\{\hat{\phi}_1, \ldots, \hat{\phi}_p, \hat{\theta}_1, \ldots, \hat{\theta}_q, \hat{\mu}, \hat{\sigma}^2\}$.
- The exact likelihood is cumbersome and maximization requires specialize numerical methods.
- The MLEs can be obtained with the `arima` function in `R`.

5.4.6 Which $ARMA$?

How do we know if an $ARMA$ model is appropriate and which $ARMA$ model to fit?

- After fitting an $ARMA$ model, we can examine the residuals.
- The `acf` function in `R` can be used to compute empirical autocorrelations of the residuals.
- If the residuals are autocorrelated, the model is not a good fit. Consider changing the parameters p and q of the $ARMA$ or using another model.

5.4.7 Which $ARMA$?

The `auto.arima` function in `R` estimates a range of $ARMA(p, q)$ models and selects the one with the best fit.

- `auto.arima` uses the *Akaike Information Criterion* (AIC) or the *Bayesian Information Criterion* (BIC) to select the model.
- Minimizing AIC and BIC amounts to minimizing the sum of squared residuals, with a penalty term that is related to the number of model parameters.

CHAPTER

SIX

BAYESIAN METHODS

Contents:

6.1 Bayes Theorem

6.1.1 Sample Space

Consider a random variable X.

- The set of all possible outcomes of X is referred to as the *sample space*.
- We will denote the sample space as $\mathcal{S}$.
- Each outcome, x, of the random variable X is called a member of $\mathcal{S}$.
- In notation $x \in \mathcal{S}$.

6.1.2 Subsets

A subset of $\mathcal{S}$ is a collection of outcomes.

- If A is a subset of $\mathcal{S}$, we write $A \subset \mathcal{S}$.
- If B is a subset of $\mathcal{S}$ and A is a subset of B, we write $A \subset B$.
- We also say that A is *contained in* B.
- We often refer to subsets of the sample space as *events*.
- The *empty set* is the subset with no elements and is denoted $\emptyset$.
- The empty set is an impossible event.

6.1.3 Example of Subsets

Let X be the result of a fair die roll.

- The sample space is $\{1, 2, 3, 4, 5, 6\}$.
- Let B be the event that X is even: $B = \{2, 4, 6\}$.
- Let A be the event that X is 2 or 4: $A = \{2, 4\}$.
- Clearly, $A \subset B \subset \mathcal{S}$.
- Let C be the event that X is -1.

- Clearly, $C = \emptyset$.

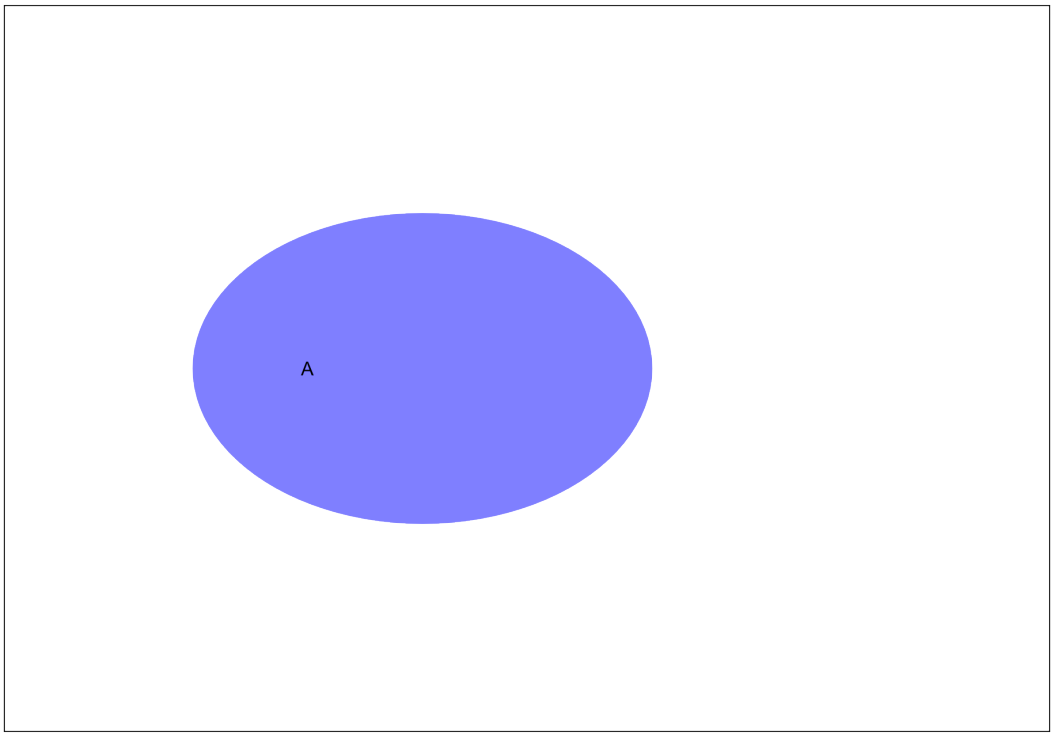
6.1.4 Union

The *union* of two sets is the set containing all outcomes that belong to A *or* B.

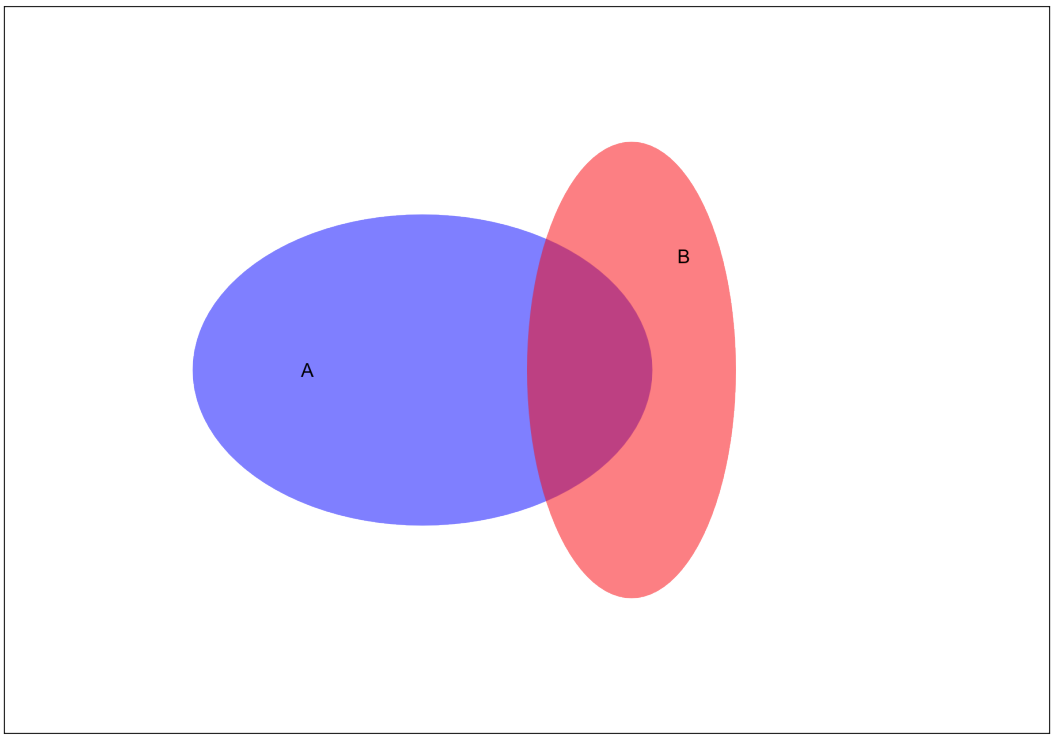
- We write the union of A and B as $A \cup B$.
- For the die roll example, if $A = \{2, 5\}$ and $B = \{2, 4, 6\}$, then $A \cup B = \{2, 4, 5, 6\}$.
- The union of many sets is written as

$$\bigcup_{i=1}^{N} A_i.$$

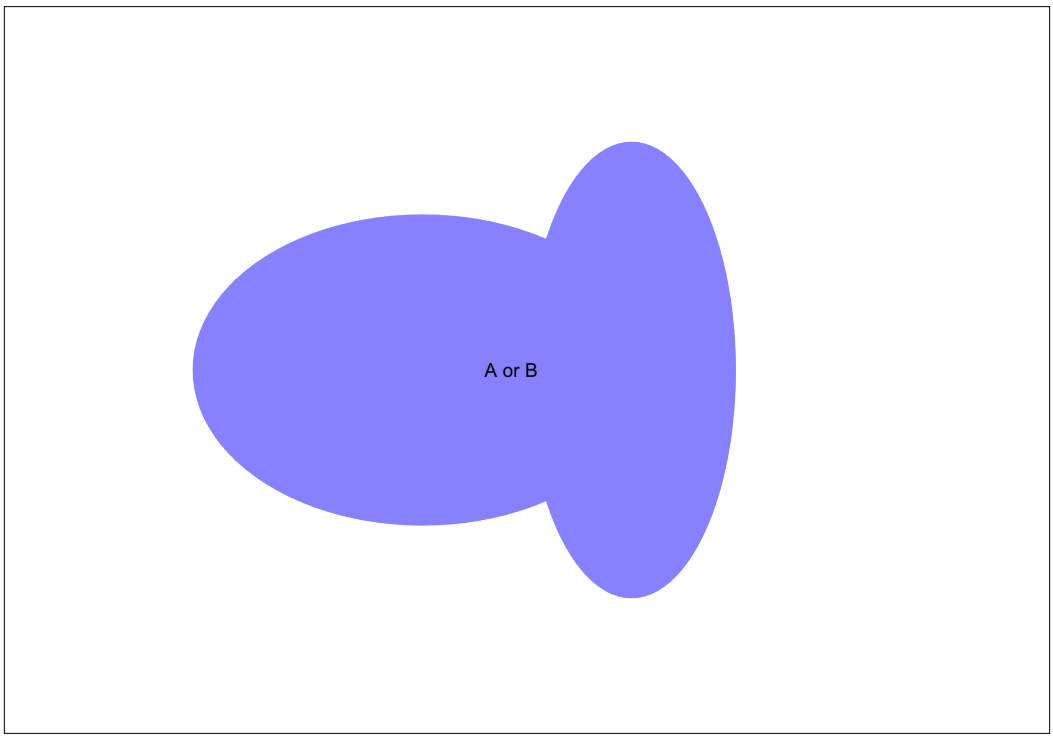
6.1.5 Union



6.1.6 Union



6.1.7 Union



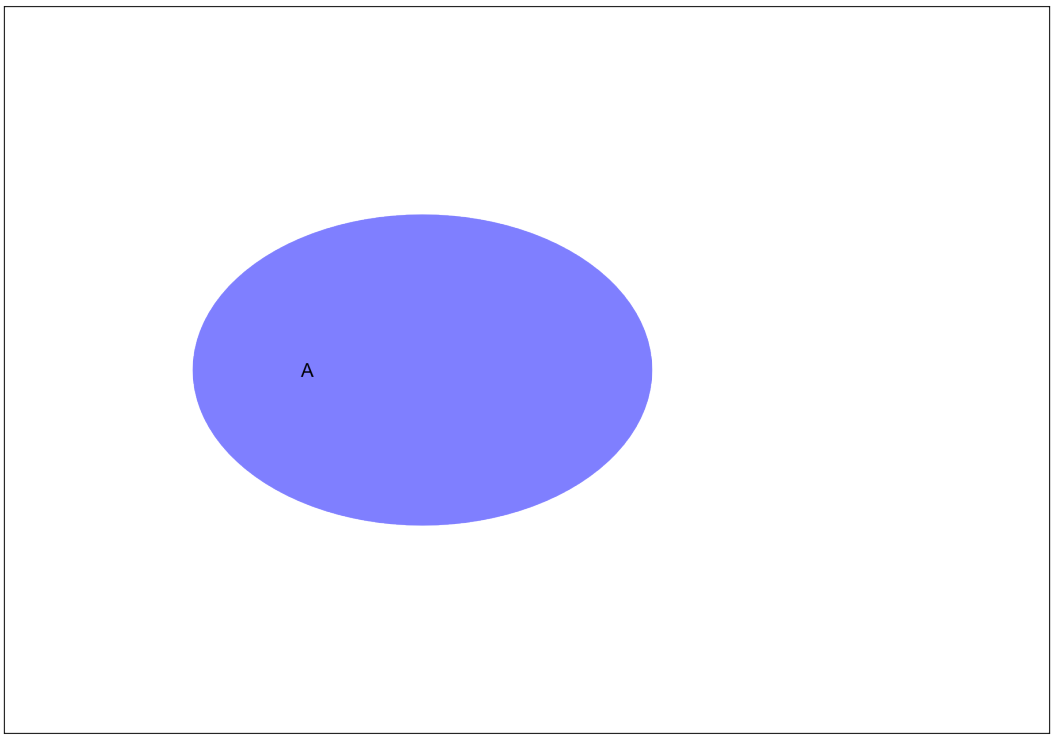
6.1.8 Intersection

The *intersection* of two sets is the set containing all outcomes that belong to A *and* B.

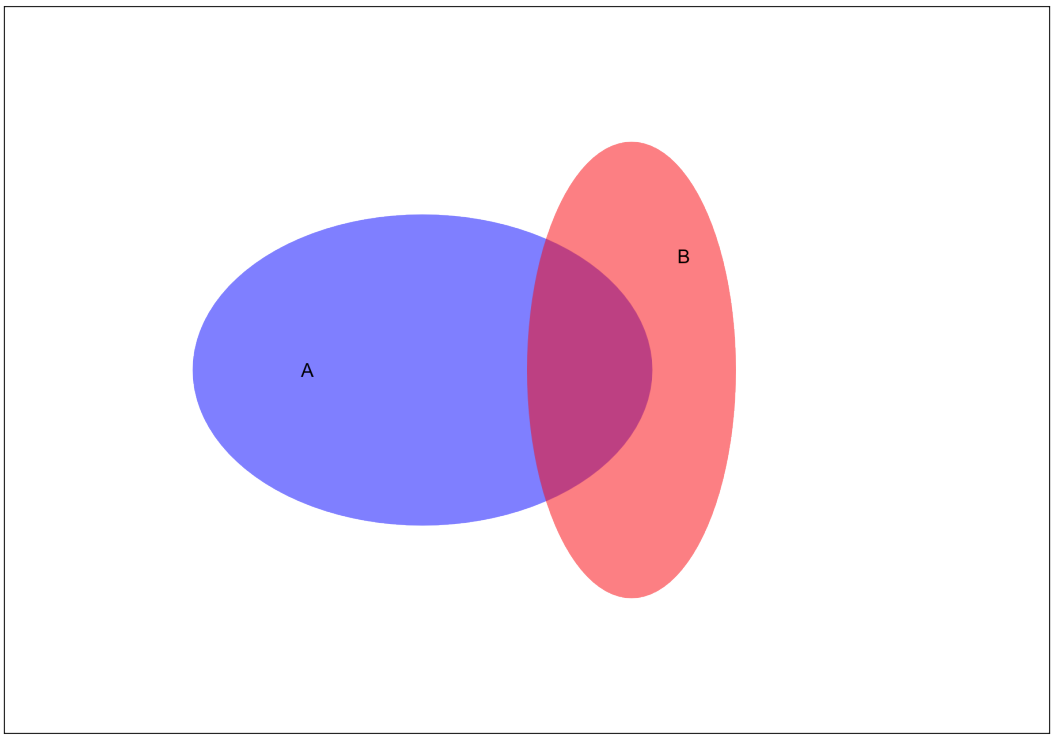
- We write the intersection of A and B as $A \cap B$.
- For the die roll example, if $A = \{2, 5\}$ and $B = \{2, 4, 6\}$, then $A \cap B = \{2\}$.
- The intersection of many sets is written as

$$\bigcap_{i=1}^{N} A_i.$$

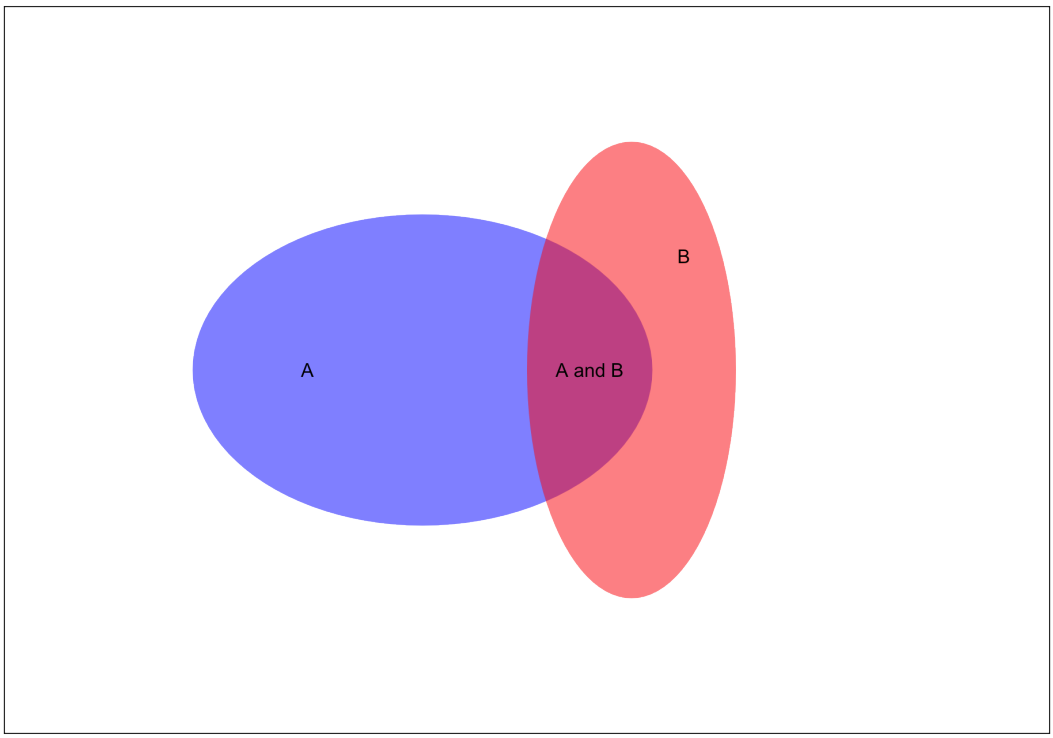
6.1.9 Intersection



6.1.10 Intersection



6.1.11 Intersection

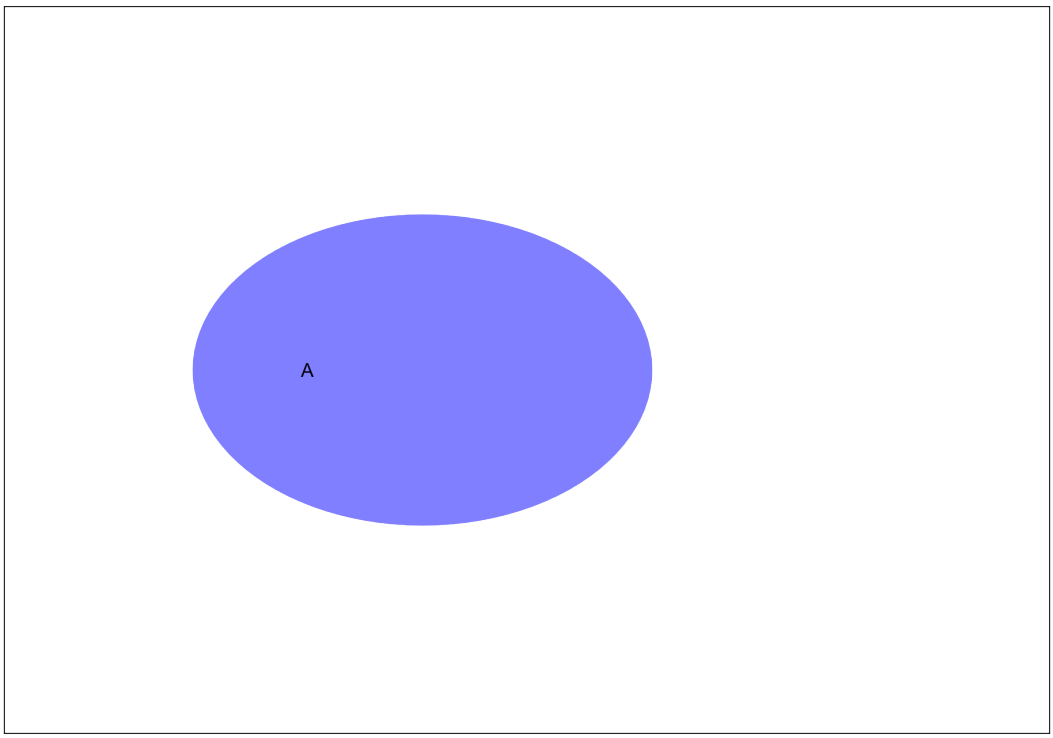


6.1.12 Complements

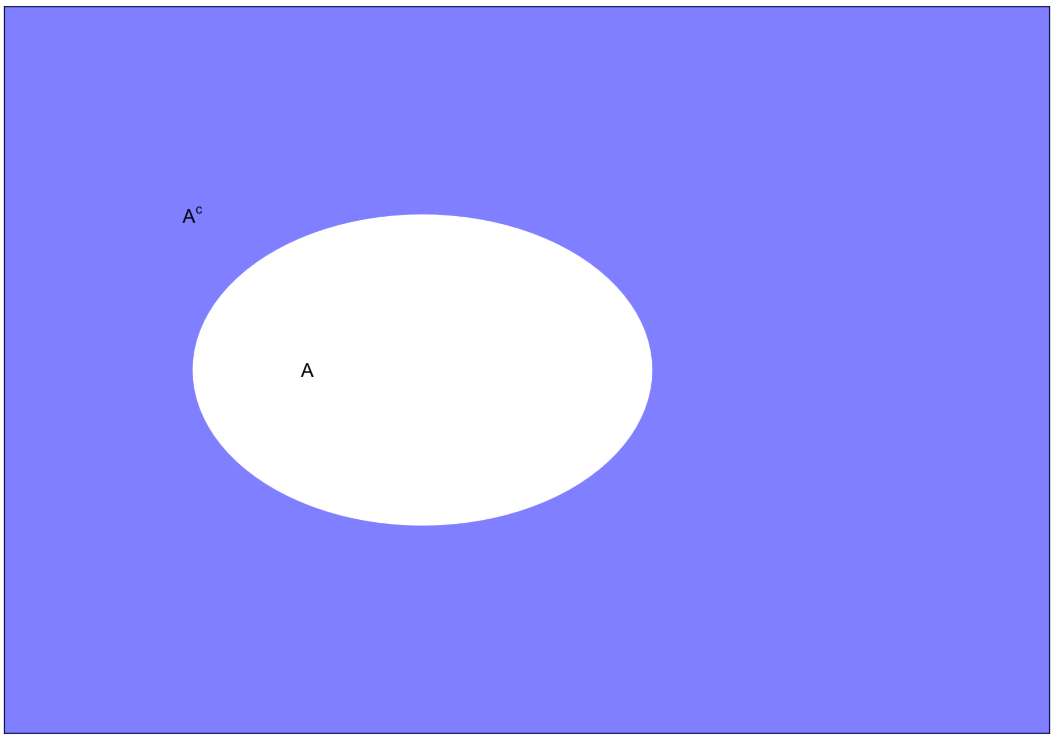
The *complement* of $A \subset \mathcal{S}$ is the subset that contains all outcomes in $\mathcal{S}$ that are not in A.

- We denote the complement by A^c.
- For the die roll example, if $A = \{2, 4, 6\}$, then $A^c = \{1, 3, 5\}$.

6.1.13 Complements



6.1.14 Complements



6.1.15 Complements

Some important properties:

$$(A^c)^c = A$$

$$\emptyset^c = \mathcal{S}$$

$$\mathcal{S}^\rfloor = \emptyset$$

$$A \cup A^c = \mathcal{S}$$

$$A \cap A^c = \emptyset.$$

6.1.16 Disjoint Events

Two events are *disjoint* or *mutually exclusive* if they have no outcomes in common.

- A and B are disjoint if $A \cap B = \emptyset$.
- By definition, any event and its complement are disjoint: $A \cap A^c = \emptyset$.
- For the die roll example, if $A = \{2, 5\}$ and $B = \{4, 6\}$, then $A \cap B = \emptyset$.

6.1.17 Probability

Given $A \subset \mathcal{S}$, denote the probability $P(X \subset A) = P(A)$.

- In the Venn diagram, $P(A)$ is the ratio of the area of A to the area of $\mathcal{S}$.
- Note that $P(\mathcal{S}) = 1$.
- Note that $P(\emptyset) = 0$.

6.1.18 Probability of Intersections

The probability that X is in A and B is:

$$P(A \cap B) = P((X \subset A) \cap (X \subset B)).$$

- For the die roll example, if $A = \{2, 5\}$ and $B = \{2, 4, 6\}$, then

$$P(A \cap B) = P(X = 2) = \frac{1}{6}.$$

6.1.19 Probability of Unions

The probability that X is in A or B is:

$$\begin{aligned} P(A \cup B) &= P((X \subset A) \cup (X \subset B)) \\ &= P(A) + P(B) - P(A \cap B). \end{aligned}$$

6.1.20 Probability of Unions

For the die roll example, if $A = \{2, 5\}$ and $B = \{2, 4, 6\}$,

$$\begin{aligned} & P(A) + P(B) - P(A \cap B) \\ &= P(\{2, 5\}) + P(\{2, 4, 6\}) - P(\{2\}) \\ &= \frac{2}{6} + \frac{3}{6} - \frac{1}{6} \\ &= \frac{4}{6} \\ &= P(\{2, 4, 5, 6\}) \\ &= P(A \cup B). \end{aligned}$$

6.1.21 Probability of Unions

If $A \cap B = \emptyset$,

$$P(A \cup B) = P(A) + P(B),$$

since $P(\emptyset) = 0$.

6.1.22 Probability of Unions

For the die roll example, if $A = \{2, 5\}$ and $B = \{4, 6\}$,

$$\begin{aligned} & P(A) + P(B) - P(A \cap B) \\ & = P(\{2,5\}) + P(\{4,6\}) - P(\emptyset) \\ & = \frac{2}{6} + \frac{2}{6} - 0 \\ & = \frac{4}{6} \\ & = P(\{2,4,5,6\}) \\ & = P(A \cup B). \end{aligned}$$

6.1.23 Conditional Probability

Suppose we know that event B has occurred - that is, one of the outcomes in the subset $B \subset \mathcal{S}$ has occurred.

- How does this alter our view of the probability of event A occurring?
- Denote the probability of A, conditional on B occurring, as $P(A|B)$.
- If $A \cap B = \emptyset$, we know $P(A|B) = 0$. Why?

6.1.24 Conditional Probability

If $A \cap B \neq \emptyset$

$$P(A|B) = \frac{P(A \cap B)}{P(B)}.$$

- $P(A|B)$ is the ratio of the area of $A \cap B$ to the area of B.
- That is, we reduce the sample space from $\mathcal{S}$ to B.

6.1.25 Conditional Probability

For the die roll example, if $A = \{2, 4\}$ and $B = \{2, 4, 6\}$,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} = \frac{\frac{2}{6}}{\frac{3}{6}} = \frac{2}{3}.$$

- Intuitively, if we know that 2,4 or 6 occurs, then the probability that a 2 or 4 occurs should be $\frac{2}{3}$.

6.1.26 Bayes' Theorem

From the definition of conditional probability,

$$P(A|B) = \frac{P(A \cap B)}{P(B)} \quad \Rightarrow \quad P(A \cap B) = P(A|B)P(B). \tag{6.1}$$

Likewise

$$P(B|A) = \frac{P(A \cap B)}{P(A)} \quad \Rightarrow \quad P(A \cap B) = P(B|A)P(A). \tag{6.2}$$

6.1.27 Bayes' Theorem

Equating (6.1) and (6.2)

$$P(A|B)P(B) = P(B|A)P(A). \tag{6.3}$$

Rearranging (6.3) gives *Bayes' Theorem*:

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)}.$$

6.1.28 Bayes' Theorem

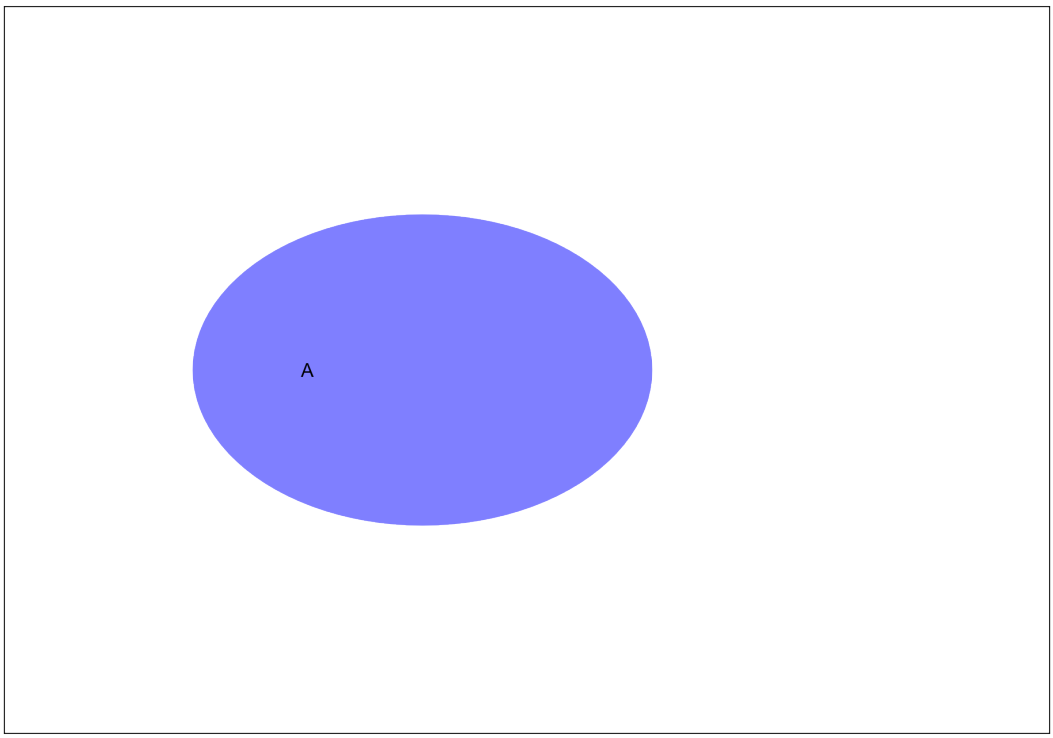
For the die roll example, if $A = \{2, 4\}$ and $B = \{2, 4, 6\}$, we already know that

- $P(A) = \frac{2}{6}$.
- $P(B) = \frac{3}{6}$.
- $P(A|B) = \frac{2}{3}$.

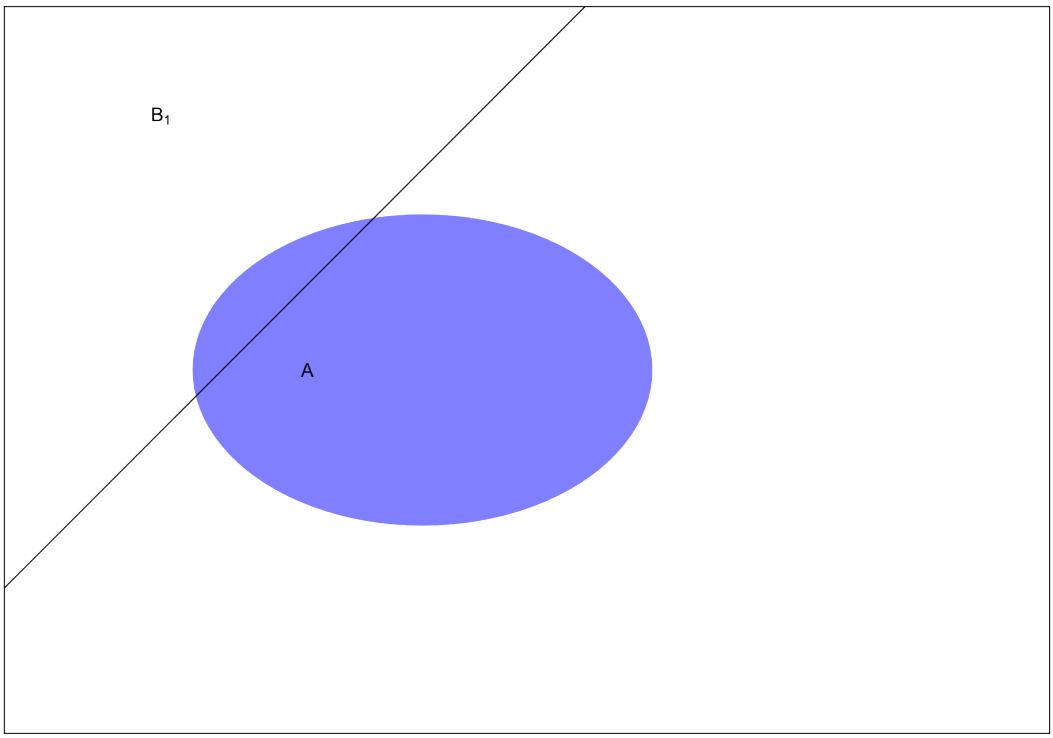
Thus,

$$P(B|A) = \frac{P(A|B)P(B)}{P(A)} = \frac{\frac{2}{3} \times \frac{3}{6}}{\frac{2}{6}} = \frac{\frac{2}{6}}{\frac{2}{6}} = 1.$$

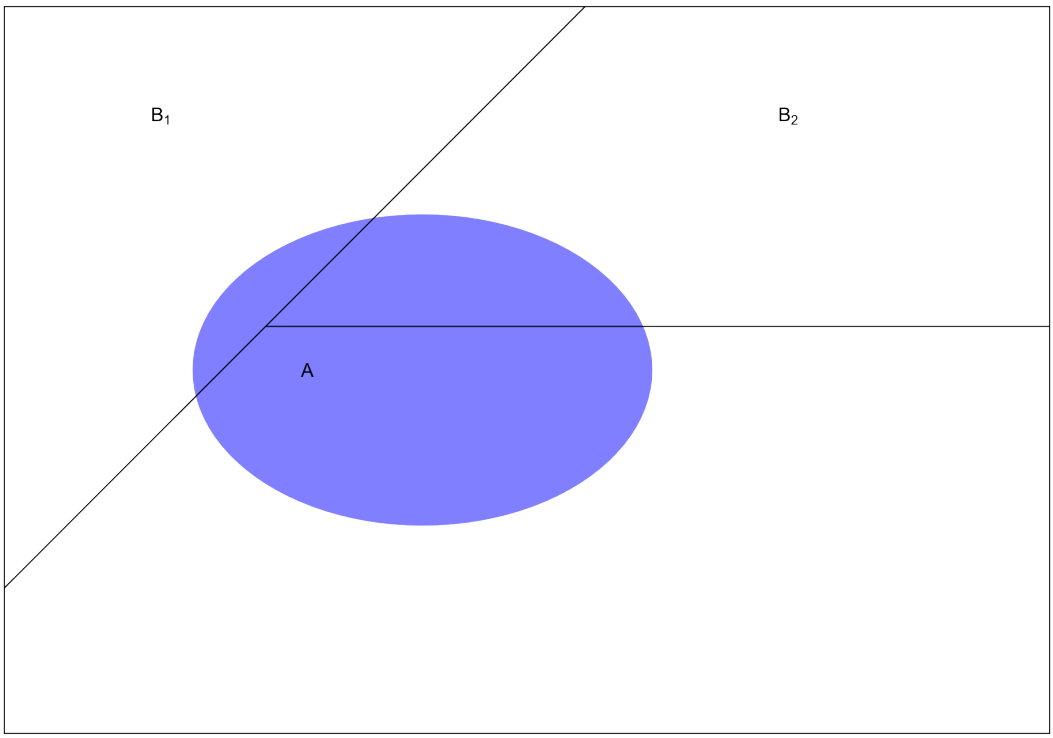
6.1.29 Partitions



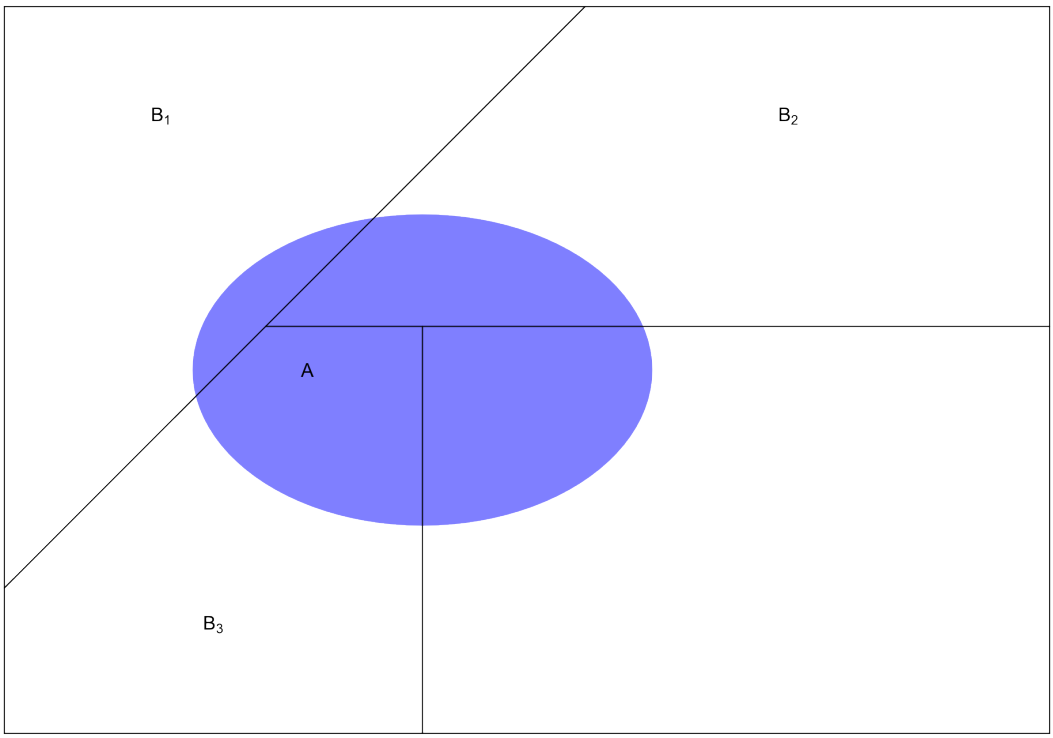
6.1.30 Partitions



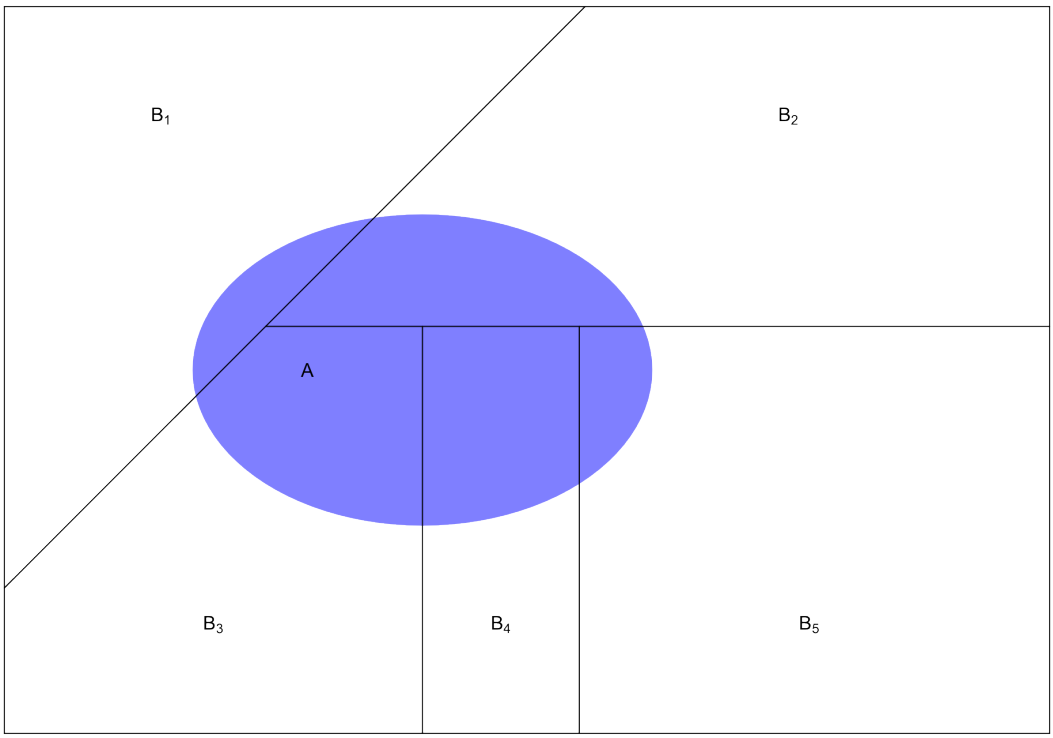
6.1.31 Partitions



6.1.32 Partitions



6.1.33 Partitions



6.1.34 Partitions

Let $B_1, B_2, \ldots, B_K$ be K subsets of $\mathcal{S}$: $B_i \subset \mathcal{S}$ for $i = 1, \ldots, K$.

- $\{B_i\}_{i=1}^K$ is a *partition* of $\mathcal{S}$ if

$$B_1 \cup B_2 \cup \cdots \cup B_K = \mathcal{S}$$

$$B_i \cap B_j = \emptyset \quad \text{for} i \neq j.$$

- Note that $A = (A \cap B_1) \cup \cdots \cup (A \cap B_K)$.

6.1.35 Partitions

Since $(A \cap B_i) \cap (A \cap B_j) = \emptyset$ for $i \neq j$,

$$P(A) = P\left((A \cap B_1) \cup \cdots \cup (A \cap B_K)\right)$$

$$= P(A \cap B_1) + \cdots + P(A \cap B_K)$$

$$= P(A|B_1)P(B_1) + \cdots + P(A|B_K)P(B_K).$$

6.1.36 Bayes' Theorem Extended

Given a partition $B_1, \ldots, B_K$ of $\mathcal{S}$, we can applied Bayes' Theorem to each subset of the partition:

$$P(B_j|A) = \frac{P(A|B_j)P(B_j)}{P(A)}$$

$$= \frac{P(A|B_j)P(B_j)}{P(A|B_1)P(B_1) + \cdots + P(A|B_K)P(B_K)}.$$

6.2 Prior and Posterior Distributions

6.2.1 Prior Distribution

Bayes' Theorem can be used to model the distribution of parameters.

- Recall that the likelihood of data $\mathbf{y}$ can be expressed as $f(\mathbf{y}|\theta)$.
- θ is a vector of parameters.
- In reality, we think of θ as a set of unknown values that are *not random*.
- However, we treat θ as random because of our lack of knowledge.
- That is, our lack of knowledge induces a distribution over θ.

6.2.2 Prior Distribution and Likelihood

The *prior distribution* $\pi(\theta)$ expresses our beliefs about θ prior to observing data $\mathbf{y}$.

- $\pi(\theta)$ is different from the likelihood: $f(\mathbf{y}|\theta)$.
- $\pi(\theta)$ is loosely interpreted as the probability of θ occurring before we observe data.
- $f(\mathbf{y}|\theta)$ is loosely interpreted as the probability of the data occurring, given a specific value of the parameter vector θ.

6.2.3 Joint Density

The joint density of $\mathbf{y}$ and θ is

$$f(\mathbf{y}, \theta) = f(\mathbf{y}|\theta)\pi(\theta).$$

- This is analogous to the relationship we previously derived:

$$P(A \cap B) = P(A|B)P(B).$$

6.2.4 Marginal Density

The marginal density of $\mathbf{y}$ is

$$f(\mathbf{y}) = \int f(\mathbf{y}, \theta) d\theta = \int f(\mathbf{y}|\theta)\pi(\theta) d\theta.$$

6.2.5 Marginal Density

- This is analogous to the relationship we previously derived:

$$\begin{aligned} P(A) &= P\left((A \cap B_1) \cup \cdots \cup (A \cap B_K)\right) \\ &= P(A \cap B_1) + \cdots + P(A \cap B_K) \\ &= P(A|B_1)P(B_1) + \cdots + P(A|B_K)P(B_K) \\ &= \sum_{i=1}^{K} P(A|B_i)P(B_i), \end{aligned}$$

for a partition $\{B_i\}_{i=1}^{K}$.

6.2.6 Posterior Distribution

According to Bayes' Theorem,

$$\begin{aligned} \pi(\theta|\mathbf{y}) &= \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} \\ &= \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int f(\mathbf{y}|\theta)\pi(\theta)d\theta}. \end{aligned}$$

- $\pi(\theta|\mathbf{y})$ is referred to as the *posterior distribution* of θ.
- $\pi(\theta|\mathbf{y})$ is loosely interpreted as the probability of θ after observing $\mathbf{y}$.

6.2.7 Bayesian Updating

Bayesian analysis is a method to use data to update our beliefs about θ.

- We begin with a prior distribution $\pi(\theta)$ which captures our views about the likelihood of θ taking particular values.
- We specify a model for the probability density of the data, given θ: $f(\mathbf{y}|\theta)$.
- We use the likelihood to update our beliefs about θ:

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{\int f(\mathbf{y}|\theta)\pi(\theta)d\theta}.$$

- If the data are very informative, $\pi(\theta|\mathbf{y})$ can be quite different from $\pi(\theta)$.

6.2.8 A Note on Proportionality

Suppose

$$w = ax$$

$$y = bx$$

$$z = wy$$

then

$$w \propto x$$

$$y \propto x$$

$$z \propto x^2.$$

6.2.9 A Note on Proportionality

More generally, if

$$w = g_w(x)h_w(u)$$

$$y = g_y(x)h_y(u)$$

$$z = wy$$

then

$$w \propto g_w(x)$$

$$y \propto g_y(x)$$

$$z \propto g_w(x)g_y(x).$$

6.2.10 A Note on Proportionality

Since $f(\mathbf{y})$ is not a function of θ,

$$\pi(\theta|\mathbf{y}) = \frac{f(\mathbf{y}|\theta)\pi(\theta)}{f(\mathbf{y})} \propto f(\mathbf{y}|\theta)\pi(\theta).$$

- It is often easier to work with only $f(\mathbf{y}|\theta)\pi(\theta)$.

6.2.11 Conjugate Priors

Our choice of $\pi(\theta)$ and $f(\mathbf{y}|\theta)$ may not yield an analytic solution for $\pi(\theta|\mathbf{y})$.

- $\pi(\theta|\mathbf{y})$ still exists, but it must be computed numerically.
- However, when the likelihood and prior have similar forms, they result in tractable posteriors.
- A *conjugate* prior is a distribution that results in a posterior of the same family when coupled with a particular likelihood.

6.2.12 Conjugate Priors

- For example, if $f(\mathbf{y}|\theta)$ is a binomial distribution and $\pi(\theta)$ is a beta distribution, $\pi(\theta|\mathbf{y})$ will also be a beta distribution.
- Alternatively, if $f(\mathbf{y}|\theta)$ is a normal distribution and $\pi(\theta)$ is a normal distribution, $\pi(\theta|\mathbf{y})$ will also be a normal distribution.

6.2.13 Normal Example

Suppose $Y_1, \ldots, Y_n \overset{i.i.d.}{\sim} \mathcal{N}(\mu, \sigma^2)$, where σ^2 is *known* and μ is *unknown*.

- Assume $\pi(\mu)$ is $\mathcal{N}(\mu_0, \sigma_0^2)$, where μ_0 and σ_0^2 are known parameters.
- We will see below that σ_0^2 provides a measure of how strong our beliefs are that $\mu = \mu_0$ prior to observing data.

6.2.14 Normal Example

The prior is

$$\pi(\mu) = \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu - \mu_0)^2\right\}$$

$$= \frac{1}{\sqrt{2\pi}\sigma_0} \exp\left\{-\frac{1}{2\sigma_0^2}(\mu^2 - 2\mu\mu_0 + \mu_0^2)\right\}$$

$$\propto \exp\left\{\frac{\mu\mu_0}{\sigma_0^2} - \frac{\mu^2}{2\sigma_0^2}\right\}.$$

6.2.15 Normal Example

The likelihood is

$$f(Y_1, \ldots, Y_n|\mu) = \prod_{i=1}^{n}\left[\frac{1}{\sqrt{2\pi}\sigma} \exp\left\{-\frac{1}{2\sigma^2}(Y_i - \mu)^2\right\}\right]$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(Y_i - \mu)^2\right\}$$

$$= \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\left(-2n\bar{Y}\mu + n\mu^2 - \sum_{i=1}^{n} Y_i^2\right)\right\}$$

$$\propto \exp\left\{\frac{n\bar{Y}\mu}{\sigma^2} - \frac{n\mu^2}{2\sigma^2}\right\}$$

6.2.16 Normal Example

The posterior is

$$\begin{aligned}\pi(\mu|Y_1,\ldots,Y_n) &\propto f(Y_1,\ldots,Y_n|\mu)\pi(\mu)\\ &\propto \exp\left\{\frac{n\bar{Y}\mu}{\sigma^2}-\frac{n\mu^2}{2\sigma^2}\right\}\exp\left\{\frac{\mu\mu_0}{\sigma_0^2}-\frac{\mu^2}{2\sigma_0^2}\right\}\\ &= \exp\left\{\left(\frac{n\bar{Y}}{\sigma^2}+\frac{\mu_o}{\sigma_0^2}\right)\mu-\left(\frac{n}{2\sigma^2}+\frac{1}{2\sigma_0^2}\right)\mu^2\right\}\\ &= \exp\left\{A\mu-\frac{B}{2}\mu^2\right\}\\ &= \exp\left\{-\frac{B}{2}\left(\mu^2-\frac{2A}{B}\mu\right)\right\}\end{aligned}$$

6.2.17 Normal Example

$$\begin{aligned}&\propto \exp\left\{-\frac{B}{2}\left(\mu^2-\frac{2A}{B}\mu\right)\right\}\exp\left\{-\frac{B}{2}\left(\frac{A}{B}\right)^2\right\}\\ &= \exp\left\{-\frac{B}{2}\left(\mu^2-\frac{2A}{B}\mu+\left(\frac{A}{B}\right)^2\right)\right\}\\ &= \exp\left\{-\frac{B}{2}\left(\mu-\frac{A}{B}\right)^2\right\}.\end{aligned}$$

6.2.18 Normal Example

We see that $\pi(\mu|Y_1,\ldots,Y_n)$ is $\mathcal{N}\left(\frac{A}{B},\frac{1}{B}\right)$ where

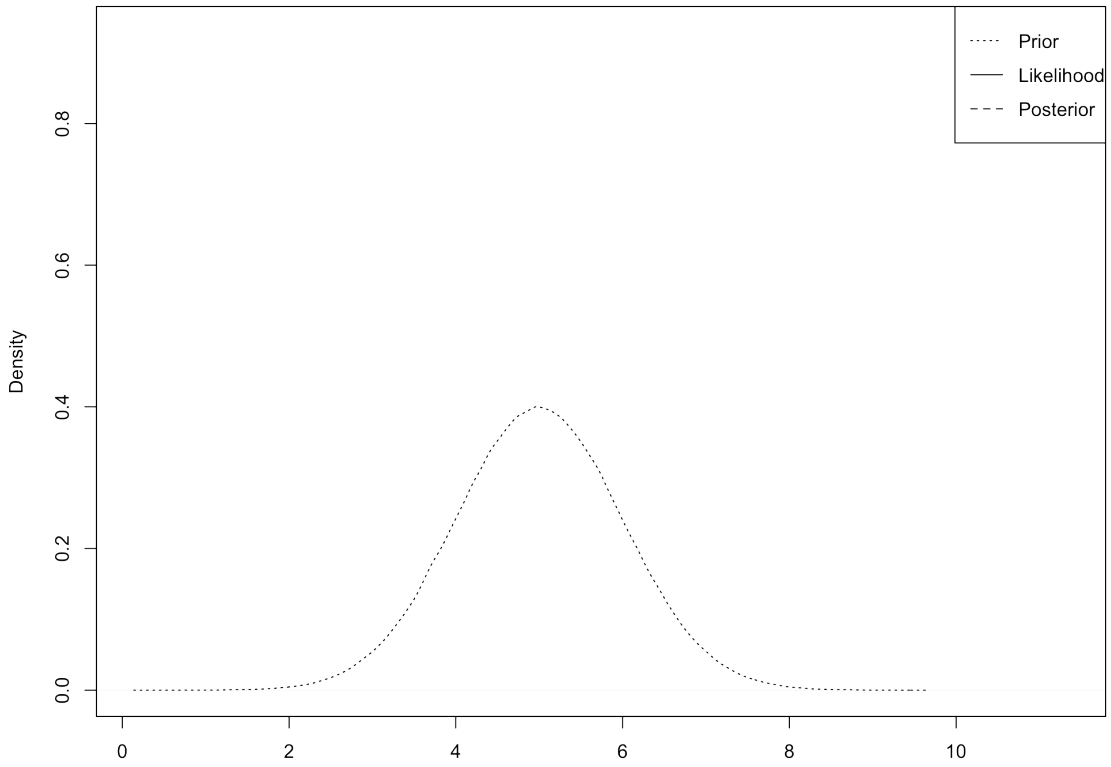
$$E[\mu|Y_1,\ldots,Y_n]=\frac{A}{B}=\frac{\frac{n\bar{Y}}{\sigma^2}+\frac{\mu_0}{\sigma_0^2}}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}$$

$$Var(\mu|Y_1,\ldots,Y_n)=\frac{1}{B}=\frac{1}{\frac{n}{\sigma^2}+\frac{1}{\sigma_0^2}}.$$

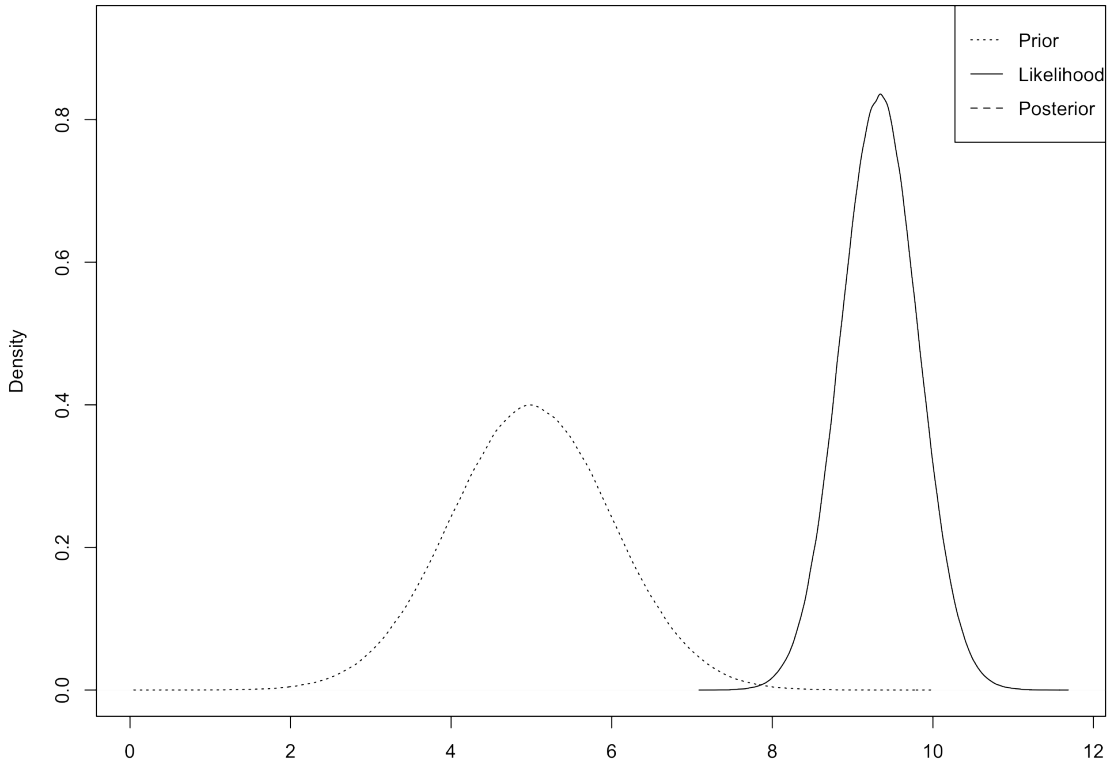
6.2.19 Normal Example

- If σ_0^2 is very small relative to σ^2/n, $E[\mu|Y_1,\ldots,Y_n]\approx\mu_0$ and $Var(\mu|Y_1,\ldots,Y_n)\approx\sigma_0^2$.
- In this case, the prior is very precise and contains a lot of information - the data doesn't add much to prior knowledge.
- If σ^2/n is very small relative to σ_0^2, $E[\mu|Y_1,\ldots,Y_n]\approx\bar{Y}$ and $Var(\mu|Y_1,\ldots,Y_n)\approx\frac{\sigma^2}{n}$.
- In this case, the prior is very imprecise and contains very little information - the data is very informative and adds a lot to prior knowledge.

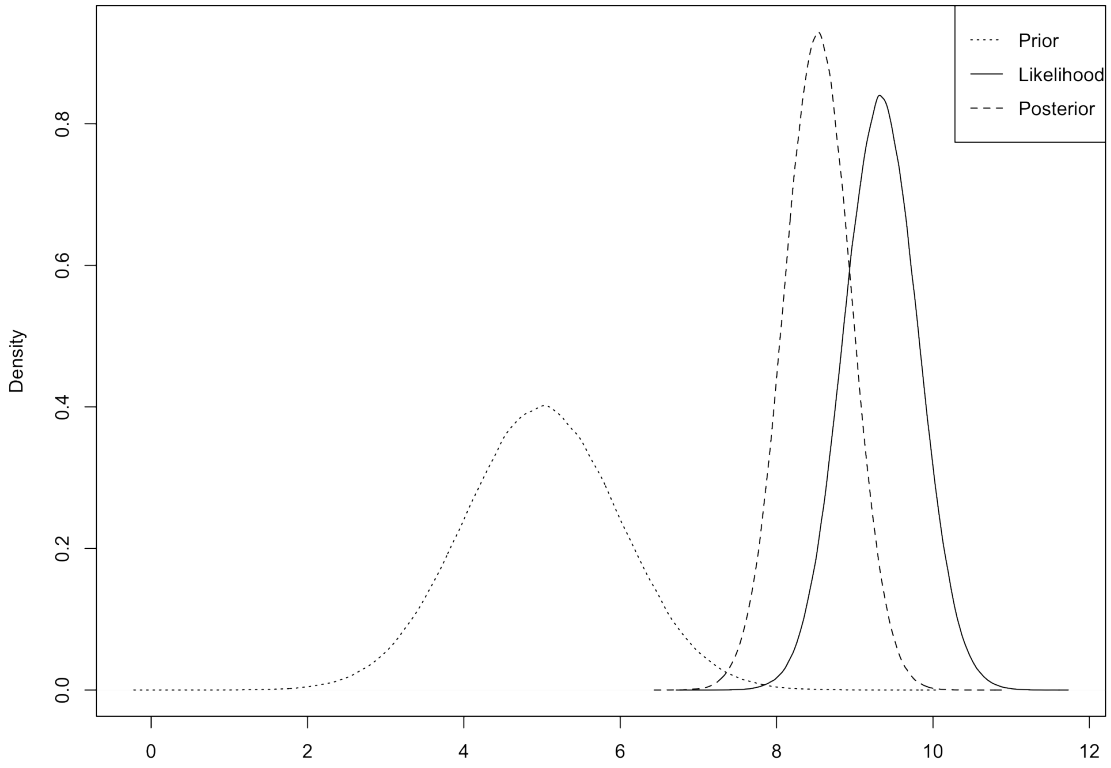
6.2.20 Moderate Prior



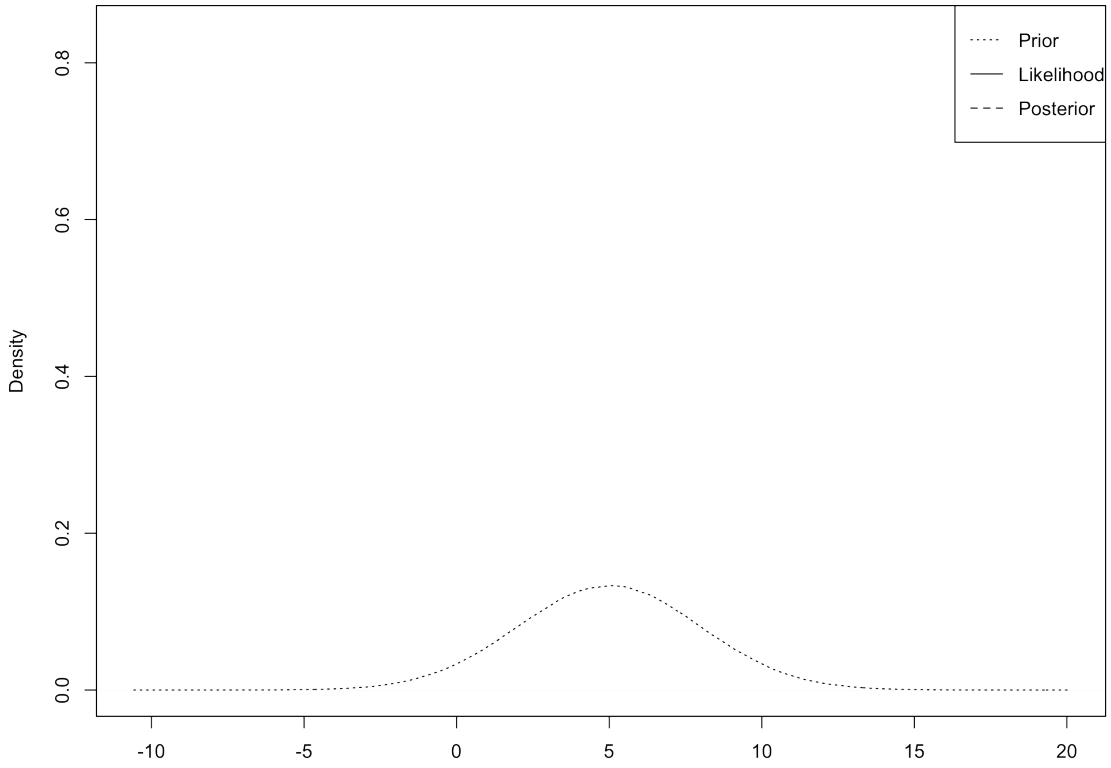
6.2.21 Moderate Prior



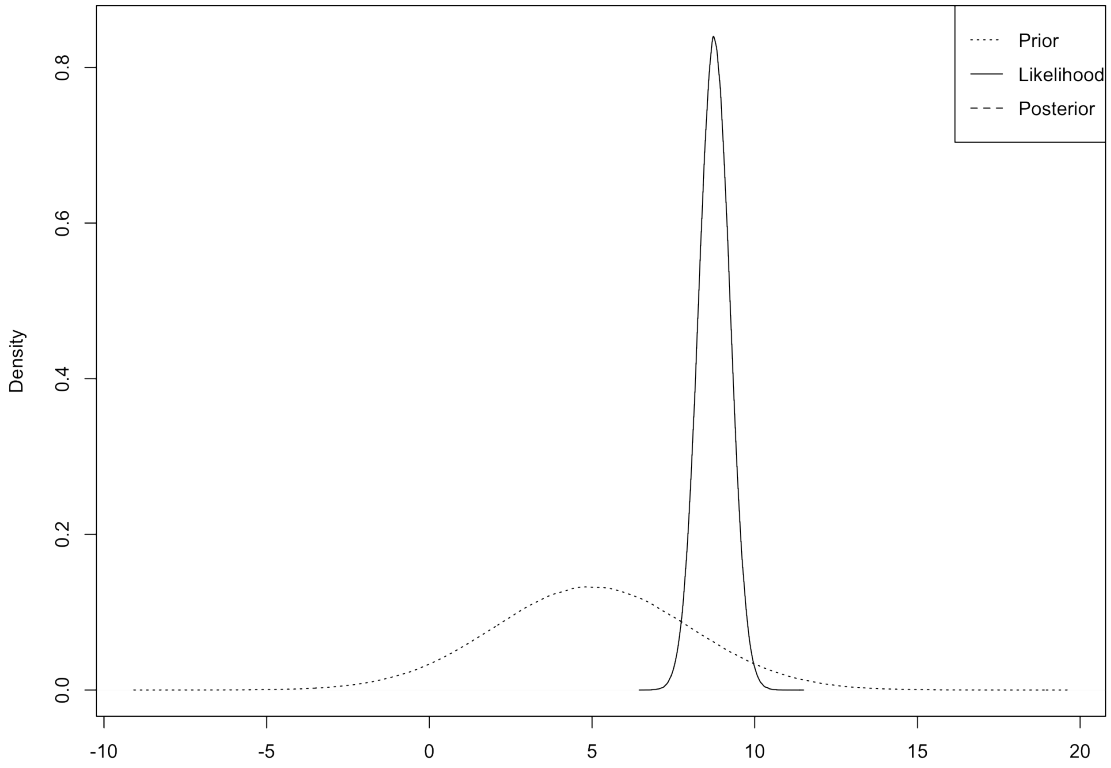
6.2.22 Moderate Prior



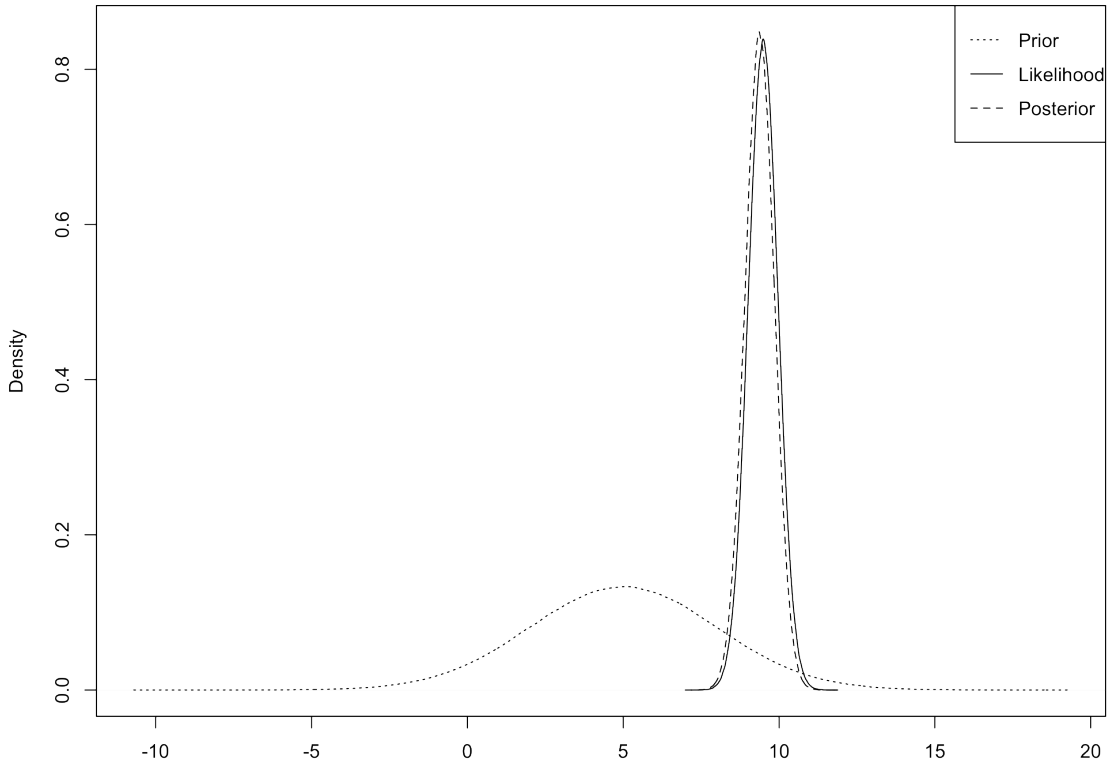
6.2.23 Uninformative Prior



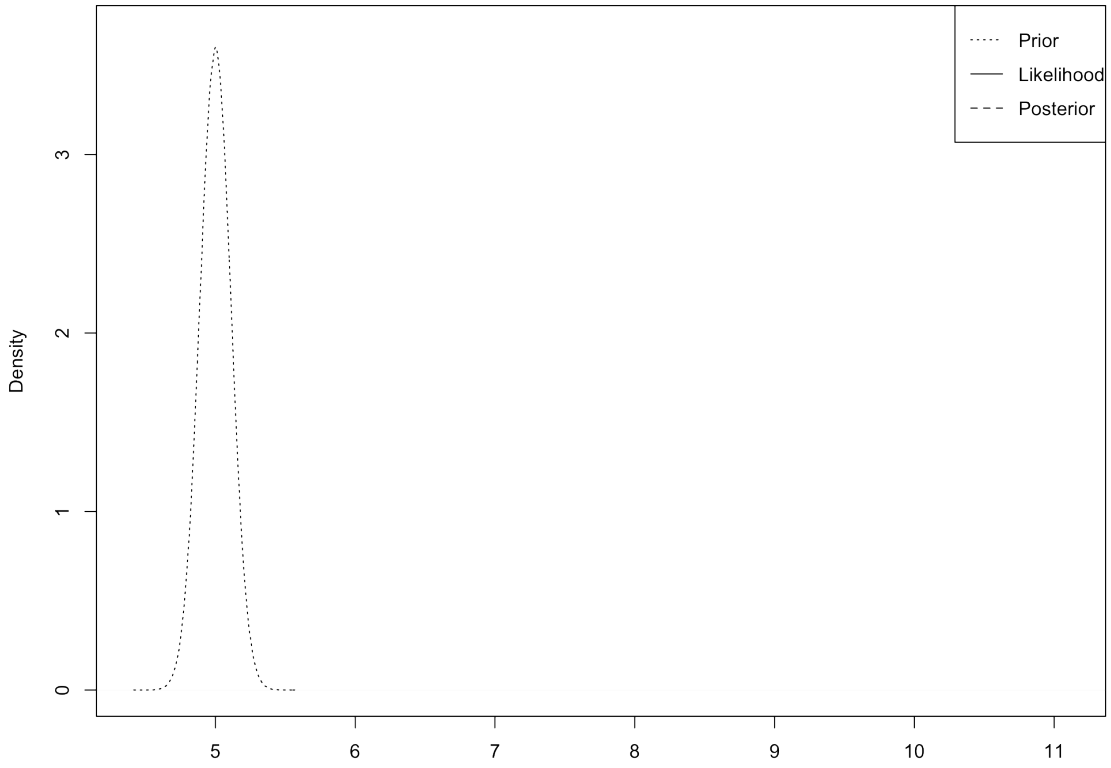
6.2.24 Uninformative Prior



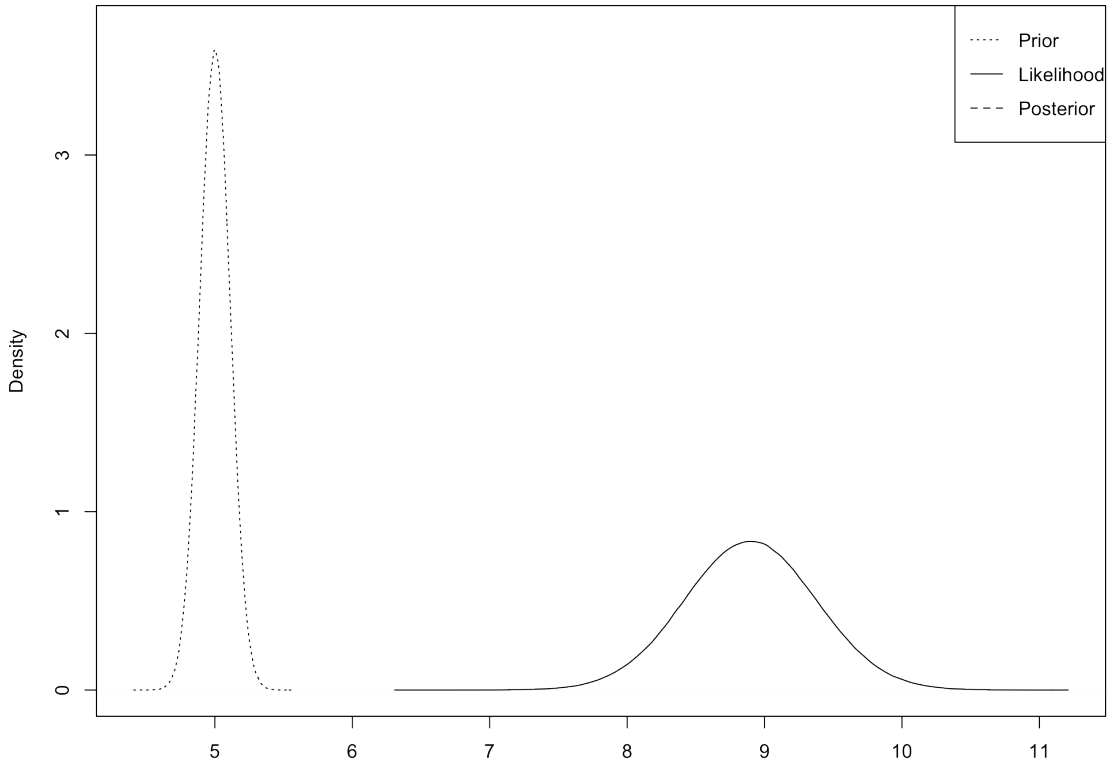
6.2.25 Uninformative Prior



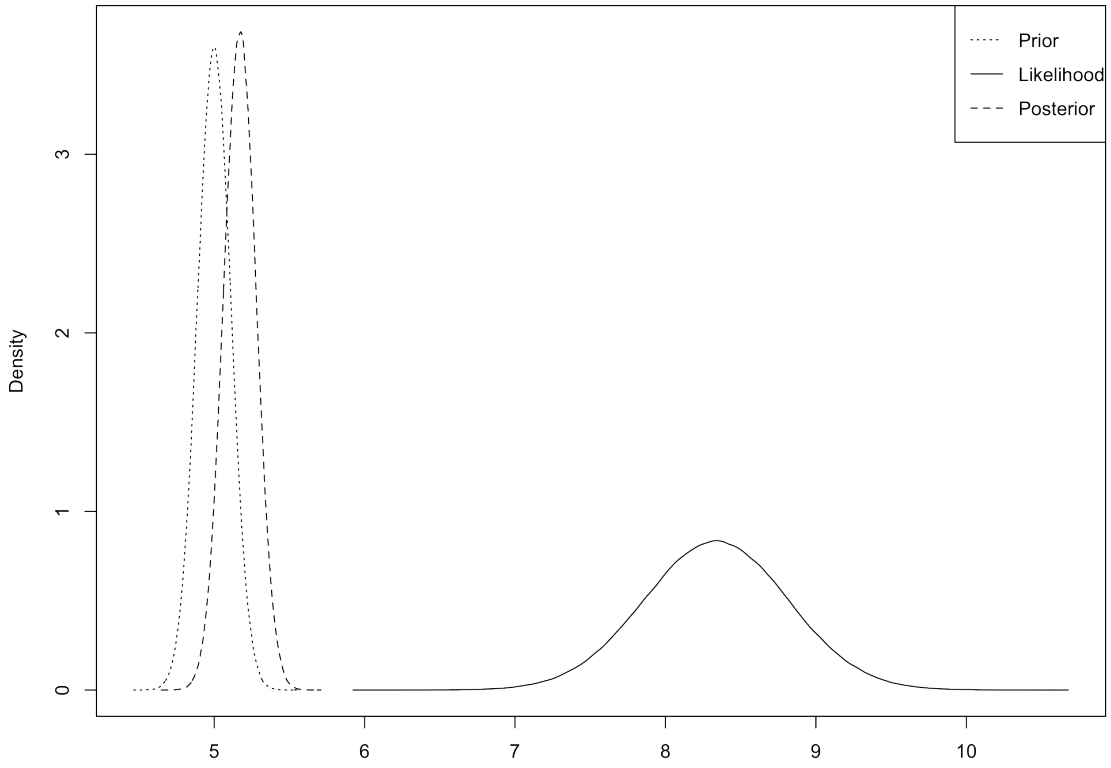
6.2.26 Informative Prior



6.2.27 Informative Prior



6.2.28 Informative Prior



6.2.29 Bayesian Parameter Estimates

The most common Bayesian parameter estimates are

- The mean of the posterior distribution.
- The mode of the posterior distribution.
- The median of the posterior distribution.
- For large n, the the mode is approximately equal to the MLE.

6.2.30 Frequentist Confidence Intervals

When constructing typical confidence intervals:

- Parameters are viewed as fixed and data as random.
- The interval is random because the data is random.
- We interpret the interval as containing the true parameter with some probability *before the data are observed*.
- Once the data are observed, the computed interval either contains or does not contain the true parameter.
- We interpret a 95% confidence interval in the following way: if we could draw 100 samples similar to the one we have, roughly 95 of the associated confidence intervals should contain the true parameter.

6.2.31 Bayesian Credible Intervals

Bayesian credible intervals are the Bayesian equivalent to frequentist confidence intervals.

- In the Bayesian paradigm, the parameters are viewed as random while the data are fixed.
- An interval based on the posterior distribution has a natural interpration as a probability of containing the true parameter, *even after the data have been observed*.

6.2.32 Equal-tails Credible Interval

The most basic $1 - \alpha$ credible interval is formed by computing the $\alpha/2$ and $1 - \alpha/2$ quantiles of the posterior distribution.

- For example, suppose $\alpha = 0.05$: you want to compute a 95% credible interval.
- Determine the 0.025 and 0.975 quantiles.
- These are the values corresponding to 2.5% of the distribution in the lower tail and 2.5% of the distribution in the upper tail.

6.2.33 Equal-tails Credible Interval