Bayes rule; Robustness of the mean

September 18, 2013

A rigorous argument for likelihood-based inference

If we're going to use likelihood to treat an unknown parameter as a random variable after we've seen the data, we should treat it as random before we see the data also. Its distribution before we see the data is called the **prior** distribution.

A rigorous argument for likelihood-based inference

If we're going to use likelihood to treat an unknown parameter as a random variable after we've seen the data, we should treat it as random before we see the data also. Its distribution before we see the data is called the **prior** distribution.

The model is in the form of a conditional probability distribution for the data x given the parameter β, $p(y \mid \beta)$. Applying the standard rule that the joint probability density of y and z is the marginal density for z, $q(z)$, times the conditional density $p(y \mid z)$, the joint pdf for the data and the parameter before we see the data is $p(y \mid \beta)\pi(\beta)$, where $\pi(\beta)$ is the prior pdf for the parameter.

A rigorous argument for likelihood-based inference

If we're going to use likelihood to treat an unknown parameter as a random variable after we've seen the data, we should treat it as random before we see the data also. Its distribution before we see the data is called the **prior** distribution.

The model is in the form of a conditional probability distribution for the data x given the parameter β, $p(y \mid \beta)$. Applying the standard rule that the joint probability density of y and z is the marginal density for z, $q(z)$, times the conditional density $p(y \mid z)$, the joint pdf for the data and the parameter before we see the data is $p(y \mid \beta)\pi(\beta)$, where $\pi(\beta)$ is the prior pdf for the parameter.

Bayes' rule

Now we apply two more rules about joint and conditional probability densities:

- The pdf of $y \mid x$ is their joint pdf divided by the marginal pdf of x.
- The marginal pdf of y is the integral over x of the joint pdf of y and x.

Putting these together gives us **Bayes' rule**:

$$r(\beta \mid y) = \frac{p(y \mid \beta)\pi(\beta)}{\int p(y \mid \beta)\pi(\beta)d\beta}.$$

$r(\beta \mid x)$ is called the **posterior** pdf.

But $r(\beta \mid x)$ is not the likelihood!

While this is true in general, in our mortgage-default example the parameter, p might have a uniform prior over $(0, 1)$. If it did, we would have $\pi(\beta) \equiv 1$, in which case $r(p \mid x)$ is indeed the likelihood, scaled to integrate to one.

But $r(\beta \mid x)$ is not the likelihood!

While this is true in general, in our mortgage-default example the parameter, p might have a uniform prior over $(0, 1)$. If it did, we would have $\pi(\beta) \equiv 1$, in which case $r(p \mid x)$ is indeed the likelihood, scaled to integrate to one.

This is generally true: the posterior pdf is the same as the normalized likelihood when $\pi(\beta)$ is constant. It might seem natural to say that a uniform prior over (0,1) in this example is reasonable if before we see the data we "have no idea" what p should be. (It's not so clear that this really is "natural", but let's not worry about that now.)

One way to put this is to say the likelihood is proportional to the posterior pdf when we have a "flat prior", which sometimes is taken to represent "ignorance".

A flat prior on the whole real line?

In our normal mean example, though, $\pi(\mu)$ can't be constant, as it then couldn't integrate to one over the entire real line. Still, it could be that, for example, $\pi(\beta)$ is a normal density with very large variance, centered not too far from $\bar{x}$. Then the sample information might make the likelihood $p(x \mid \mu)$ concentrate over a short interval relative to the standard deviation of the prior, so that over the relevant range the prior pdf is nearly constant. Then the posterior pdf would be nearly the same as the likelihood normalized to integrate to one.

Reporting the likelihood for a diverse audience

A second argument for reporting the likelihood as if it were the pdf is that if readers of your work might have different prior pdf's $\pi(\beta)$, they can all construct their own posterior pdf's $r(\beta \mid x)$ using your reported likelihood. r is just the likelihood times $\pi(\beta)$, normalized to integrate to one.

Robustness

It sounds as if we should always prefer "robust" statistical procedures — what are the other ones, "feeble"?

Robustness

It sounds as if we should always prefer "robust" statistical procedures — what are the other ones, "feeble"?

Robust procedures come in to play when there is some aspect of the uncertainty we face that we care about a lot, while we are not interested in, and hence hope to avoid having to think about and carefully model, other aspects of the uncertainty.

Robustness

It sounds as if we should always prefer "robust" statistical procedures — what are the other ones, "feeble"?

Robust procedures come in to play when there is some aspect of the uncertainty we face that we care about a lot, while we are not interested in, and hence hope to avoid having to think about and carefully model, other aspects of the uncertainty.

Why single out the mean?

We have been discussing two simple examples where we were estimating $E[x_j]$, the population mean of an i.i.d. sample of random variables. If they were returns on the stock market, an argument for caring about the mean (rather than, say, the median or mode) is that if we bought a market index fund and held it for a long time, we might expect the average return over that long time to be $E[x_j]$, which would be true regardless of what the distribution of x_j might be, so long as the distribution had a finite mean.

Why single out the mean?

We have been discussing two simple examples where we were estimating $E[x_j]$, the population mean of an i.i.d. sample of random variables. If they were returns on the stock market, an argument for caring about the mean (rather than, say, the median or mode) is that if we bought a market index fund and held it for a long time, we might expect the average return over that long time to be $E[x_j]$, which would be true regardless of what the distribution of x_j might be, so long as the distribution had a finite mean.

Even here, though, some thought is required. Stock and Watson have a chart of the the daily percentage changes in the Dow Jones Industrial Average on their p.39. You can reproduce it for yourself from the data, which they post as `BadDayOnWallStreet_Box.xls` on their data site. (R

can read Stata data sets using the `foreign` package, or you can open an Excel file, save it as a `.csv` file, and then import it into R). Their data is the *percentage change* in the DJIA. The average percentage change does *not* cumulate over time to become the long run average rate of return. [Discuss.]

The food expenditures case

If the data were food expenditures by households with a single parent and three or more children under four, it's not so clear why the mean would be our focus. If we were doing a marketing study, trying to help grocery stores to locate based on demographics, the mean might be what interests us. But if it were a public policy study, where the concern is how much of the family budget is devoted to food, or whether expenditures are sufficient to provide good nutrition for the family, the mean is unlikely to be the only aspect of the distribution that interests us. If the distribution were fat-tailed or bimodal, that would be important to know.

Assumptions about the distribution that suggest focus on the mean

We have already noted that the likelihood function for a normal distribution with known variance σ^2 depends on $\{x_1, \ldots, x_n\}$ only through $\bar{x}$:

$$\log(p(x_1, \ldots, x_n \mid \mu, \sigma^2)) = -\frac{N}{2}\log(2\pi\sigma^2) - \frac{1}{2\sigma^2}\sum_j x_j^2 - \frac{1}{2\sigma^2}N\left(-2\bar{x}\mu + \mu^2\right).$$

Only the last term depends on μ, and it involves $\{x_1, \ldots, x_n\}$ only through $\bar{x}$. This makes $\bar{x}$ a **sufficient statistic** for the mean.

Sufficient statistics

Whenever the likelihood factors into (or the log likelihood additively breaks into) terms such that one involves the unknown parameter or parameters and some function $f(x_1, \ldots, x_n)$, while the rest of the likelihood depends on x, but not the parameter(s), $f(\vec{x})$ is a sufficient statistic for the parameter. (We're using the notation $\vec{x}$ to refer to the vector $x = \{x_1, \ldots, x_n\}$. We may also sometimes simply write x for $\vec{x}$, when the meaning is clear from context.)

Sufficient statistics

Whenever the likelihood factors into (or the log likelihood additively breaks into) terms such that one involves the unknown parameter or parameters and some function $f(x_1, \ldots, x_n)$, while the rest of the likelihood depends on x, but not the parameter(s), $f(\vec{x})$ is a sufficient statistic for the parameter. (We're using the notation $\vec{x}$ to refer to the vector $x = \{x_1, \ldots, x_n\}$. We may also sometimes simply write x for $\vec{x}$, when the meaning is clear from context.)

When a sufficient statistic is available, the likelihood, and therefore the posterior pdf, depend on the data only through the sufficient statistic.

Robustness and sufficient statistics

If we are basing our inference entirely on a statistic $S(\vec{x})$, for example $S(\vec{x}) = \bar{x}$ in our first example, we are acting as if we do not expect to learn much from looking at other aspects of the sample. That is, we are acting as if the statistic we are using in our analysis is a sufficient statistic, at least approximately. Conducting our inference on the assumption that x is distributed as i.i.d. $N(0, \sigma^2)$ with known variance is in that sense a conservative assumption. It justifies the notion that $\bar{x}$ is all we should be interested in.

Asymptotic approximation

You should be familiar with the central limit theorem. A simple version of it states that

Theorem 1. *If* $\{x_1, \ldots, x_n\}$ *are i.i.d. with* $E[x_j] = \mu$ *and* $\mathrm{Var}(x_j) = \sigma^2$ *for all* j*, then*

$$\sqrt{n}(\bar{x} - \mu) \xrightarrow[n\to\infty]{\mathcal{D}} N(0, \sigma^2) \, .$$

This result does not depend on the distribution of x_j, only on the i.i.d. assumption and finite mean and variance. It implies that in "large samples" we get approximately correct distribution theory for the estimator $\bar{x}$ by using the distribution theory for the case where $x_j \sim N(\mu, \sigma^2)$, even when that is not the true distribution.

Asymptotics of likelihood

If all we observed were $\bar{x}$ and we wanted to construct a likelihood function, we could invoke the asymptotic approximation that $\bar{x} \mid \mu \sim N(\mu, \sigma^2/N)$, which implies a likelihood proportional to a $N(\bar{x}, \sigma^2/N)$ pdf, as a function of μ. This does not mean the actual likelihood has approximately this normal shape, since the actual likelihood, with non-normal x_j, would in general depend on the sample through more than just $\bar{x}$.

Asymptotics of likelihood

If all we observed were $\bar{x}$ and we wanted to construct a likelihood function, we could invoke the asymptotic approximation that $\bar{x} \mid \mu \sim N(\mu, \sigma^2/N)$, which implies a likelihood proportional to a $N(\bar{x}, \sigma^2/N)$ pdf, as a function of μ. This does not mean the actual likelihood has approximately this normal shape, since the actual likelihood, with non-normal x_j, would in general depend on the sample through more than just $\bar{x}$.

Instead, this result is about "limited information likelihood". It is the approximate likelihood under the assumption that we cannot see the whole sample, only $\bar{x}$.

What is a "large" sample?

While the central limit theorem makes no assumptions except finite mean and variance, the accuracy of the asymptotic approximation at a given sample size can be arbitrarily poor, no matter how big the sample. Using the asymptotic theory implicitly makes assumptions about the distribution, requiring that it not be so far from the normal distribution as to make the asymptotic approximation poor.

Example 1 of bad asymptotics

$$x_j = \begin{cases} 10,000 \text{ with probability } 1/10,000 \\ 0 \text{ with probability } 9,999/10000\,. \end{cases}$$

In a sample of size 1000, the most likely sample consists entirely of zeros, so $\bar{x} = 0$, $s^2 = 0$. The next most likely sample contains a single occurrence of $x_j = 10,000$, with the rest of the sample zeros. This makes $\bar{x} = 10$, $s^2 = 100,000$, and the estimated standard error of $\bar{x}$ 10. The actual values are $\mu = 1$, $\sigma^2 = 10,000$, so the true standard error of $\bar{x}$ is $\sqrt{10}$. Obviously a sample size of 1000 is not enough to make the asymptotic theory work well for this distribution, even though it does have finite mean and variance.

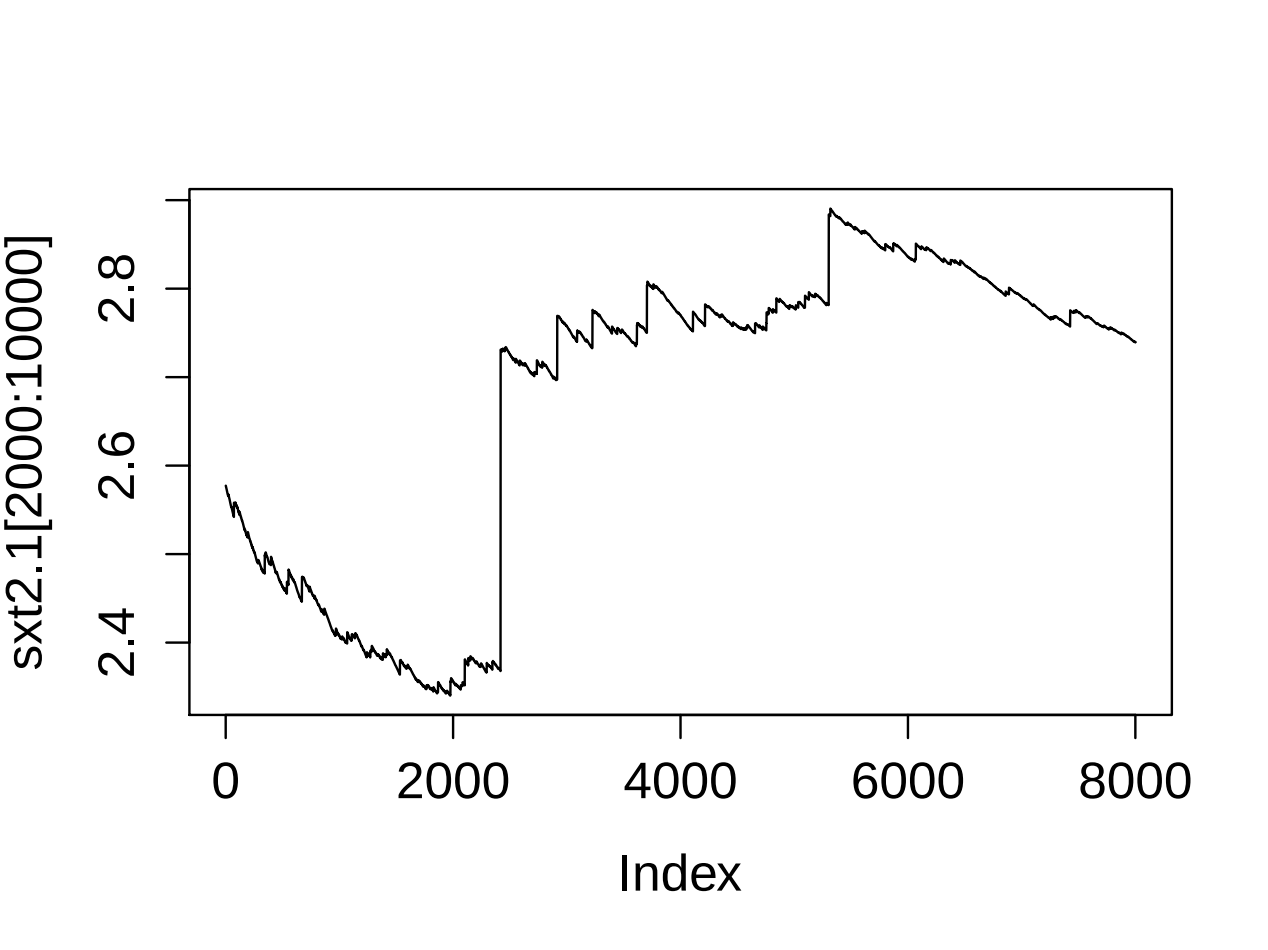
Example 2 of bad asymptotics

The t distribution with mean μ, scale parameter σ, and degrees of freedom ν has a density proportional to

$$\frac{1}{\left(1 + \frac{(x_j - \mu)^2}{\nu\sigma^2}\right)^{(\nu+1)/2}} .$$

When the degrees of freedom parameter $\nu \leq 2$, this distribution has infinite variance, and when $\nu = 1$ it does not have an expected value. (Calculus exercise: prove this.) The central limit theorem therefore does not apply. and the asymptotic normal theory for $\bar{x}$ is just not useful. If $\nu = 2+\varepsilon$, on the other hand, mean and variance of the distribution are finite. Nonetheless, if ε is very small, $\bar{x}$ and s^2 are *nearly* useless except in *extremely* large samples.

Sample standard deviations for $t(2.1)$ distribution



Efficiency loss

If we know something about the way the distribution of the data might differ from a normal distribution, this can sometimes deliver substantial increases in the accuracy of our inference.

A simple example: $x_j \sim U(\mu - 1, \mu + 1]$. The likelihood function for μ is constant between $\max\{x_1, \ldots, x_n\} - 1$ and $\min\{x_1, \ldots, x_n\} + 1$. Both ends of the interval where the likelihood is positive converge toward μ at the rate $1/N$. That is, if we want to find a limiting distribution for them as estimates of μ, we need to multiply their deviation from μ by N, not $\sqrt{N}$ as in the normal asymptotic approximation.

Inefficient inference is not "conservative"

Sometimes people speak as if using the asymptotic normal theory, which will find it more difficult to reject hypotheses and deliver longer confidence and probability intervals, is conservative.

Inefficient inference is not "conservative"

Sometimes people speak as if using the asymptotic normal theory, which will find it more difficult to reject hypotheses and deliver longer confidence and probability intervals, is conservative.

But if the object is to check the size of some important policy parameter or test an important hypothesis, there is nothing conservative about using methods that deliver greater error. Inaccurate characterization of uncertainty is not conservative.

Dangers of casual modeling

Often a main concern about the normality assumption is that the data have "fat tails" — occasional large positive or negative deviations from the center of the distribution.

Dangers of casual modeling

Often a main concern about the normality assumption is that the data have "fat tails" — occasional large positive or negative deviations from the center of the distribution.

Student t distributions have fat tails, and include the normal distribution as a special case (infinite degrees of freedom). Why not use the t distribution instead of the normal?

This actually can work well, but only if the data distribution is symmetric about μ, or if we are actually interested in the median or mode rather than the mean.

Dangers of casual modeling

Often a main concern about the normality assumption is that the data have "fat tails" — occasional large positive or negative deviations from the center of the distribution.

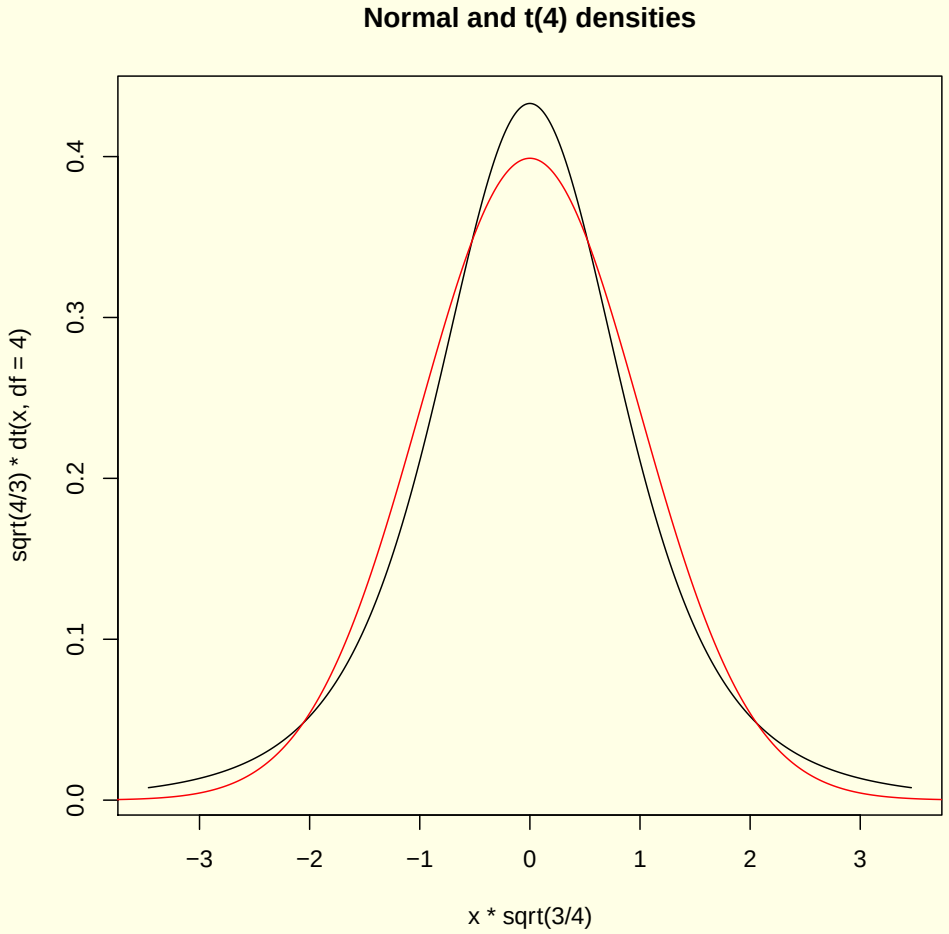
Student t distributions have fat tails, and include the normal distribution as a special case (infinite degrees of freedom). Why not use the t distribution instead of the normal?

This actually can work well, but only if the data distribution is symmetric about μ, or if we are actually interested in the median or mode rather than the mean.

Pitfalls of the t distribution with asymmetry

The t pdf has more probability weight near the center of the distribution than does a normal, and also more probability weight far out in the tails, where "outliers" occur.

Problem: the t likelihood down-weights large outliers. But if large negative (e.g.) outliers are in fact more likely than large postive outliers, downweighting both tends to increase the sample mean above $E[x_j]$. The outliers are important for determining the mean, while the t likelihood tries to exploit the relatively heavy concentration of data near μ.


Normal and t(4) densities
sqrt(4/3) * dt(x, df = 4)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
x * sqrt(3/4)

Double-exponential example

Another distribution that has fatter tails than the normal (though not as fat as the t) is the double-exponential:

$$p(x_j \mid \mu, \sigma) = \sigma^{-1} \exp\left(\frac{-|x_j - \mu|}{\sigma}\right) .$$

With σ known, its log likelihood is, up to a constant additive term,

$$\sum_{j=1}^{n} \frac{|x_j - \mu|}{\sigma} .$$

This leads to the sample median as the maximum likelihood estimator of μ.

Pitfalls of the double-exponential

If the distribution is in fact double-exponential, this will be a better estimate of μ than $\bar{x}$, but of course if the distribution is asymmetric about μ, the median and $E[x_j]$ are different, so the median will be a poor estimator of the mean. In some cases this is fine — we may actually be more interested in the median of the true distribution than in $E[x_j]$.

Making sense of all this for practice

- There is no single correct answer on these issues.

Making sense of all this for practice

- There is no single correct answer on these issues.
- One approach to robustness is to write down a model and obtain a likelihood, then check whether making the model more flexible by adding parameters makes it fit better. "Robustness via flexibility."
- Another approach is to use models that are conservative, in the sense that they imply that the estimators and other statistics we are using for our inference contain all the information about the data that is useful. In other words, ignore some information in the data and use a model that implies this is optimal. "Robustness via inefficiency."

- Each approach has dangers. Inefficiency, if it is extreme, is a bad thing. Flexibility, since it is based on our guesses about which extra parameters might be important, can, if we are mistaken, lead us to biased estimates instead of to efficiency.

Tools for checking normality: quantile-quantile plots

- `qqnorm()`: This plots the **quantiles** of the sample distribution against the corresponding quantiles of the normal distribution. For near-normally distributed data, the plot should be nearly a straight line. Fat tails show up as rapid rises in the plot at the ends.

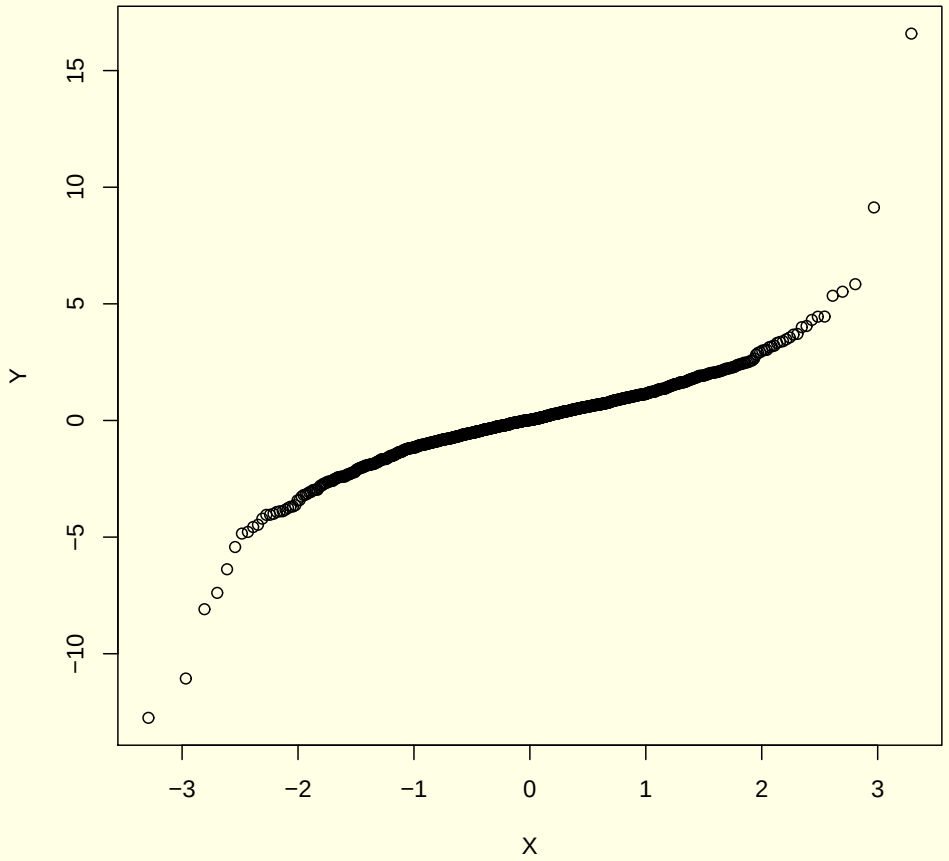
- The α-**quantile** of the distribution of a random variable x is the number γ such that $P[x \leq \gamma] = \alpha]$. It is closely related to the familiar idea of the percentile of a distribution of grades on a test.

- A qq plot has on the y axis the range of values in some data set of y values. For each of these y values, to find the corresponding x value, we find what quantile $\alpha(y)$ of the y dataset the y value represents, determine the $\alpha(y)$ quantile of a $N(0, 1)$ distribution, and plot that as the x ordinate corresponding to our y ordinate.

In case code is clearer than words to you

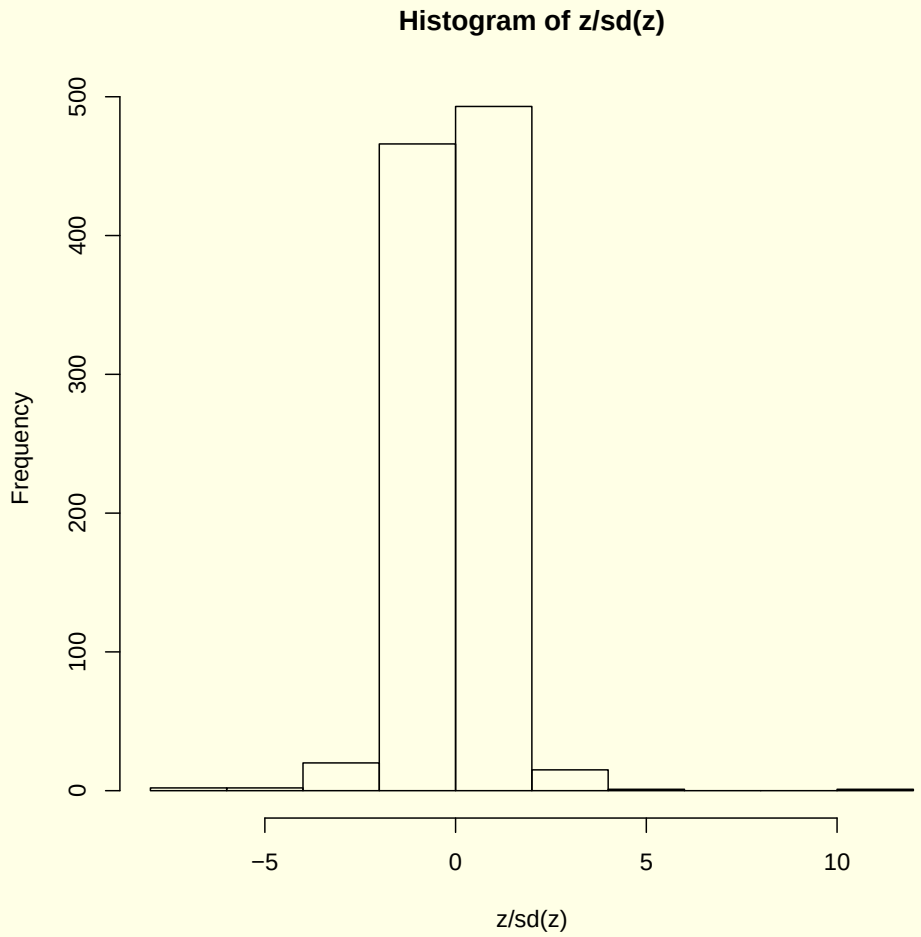
```
qq <- function(Y) {
   Y <- sort(Y)
   n <- length(Y)
   X <- vector("numeric", length(Y))
   for (i in 1:length(Y))
      X[i] <- qnorm((i-.5)/n)
   plot(X, Y)
}
```

qq plot of a $t(3)$ sample



Histograms

`hist()`: This plots a **histogram** of the data. For near-normal data, it should be bell-shaped, like a normal pdf. A normal pdf of the same variance can be plotted on the same graph, for comparison.


Histogram of z/sd(z)
Frequency
0
100
200
300
400
500
−5
0
5
10
z/sd(z)