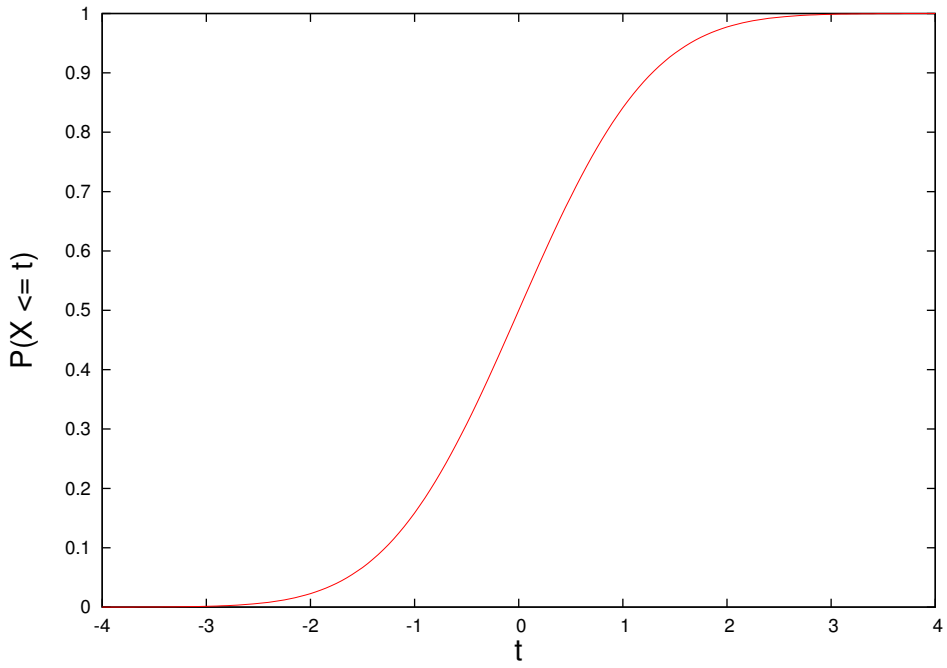
Review of basic probability and statistics

Probability: basic definitions

- A *random variable* is the outcome of a natural process that can not be predicted with certainty.
  - Examples: the maximum temperature next Tuesday in Chicago, the price of Wal-Mart stock two days from now, the result of flipping a coin, the response of a patient to a drug, the number of people who will vote for a certain candidate in a future election.
  - On the other hand, the time of sunrise next Tuesday is for all practical purposes exactly predictable from the laws of physics, and hence is not really a random variable (although technically it may be called a *degenerate* random variable).
  - There is some grayness in this definition: eventually we may be able to predict the weather or even sociological phenomena like voting patterns with extremely high precision. From a practical standpoint this is not likely to happen any time soon, so we consider a random variable to be the state of a natural process that human beings cannot currently predict with certainty.
- The set of all possible outcomes for a random variable is called the *sample space*. Corresponding to each point in the sample space is a *probability*, which is a number between 0 and 1. The sample space together with all probabilities is called the *distribution*.
- Properties of probabilities: (i) a probability is always a number between 0 and 1, (ii) the sum of probabilities for all points in the samples space is always exactly 1.
  - **Example:** If X is the result of flipping a fair coin, the sample space of X is $\{H, T\}$ (H for heads, T for tails). Either outcome has probability $1/2$, so we write $P(X = H) = 1/2$ (i.e. the probability that X is a head is $1/2$) and $P(X = T) = 1/2$. The distribution can be written $\{H \to 1/2, T \to 1/2\}$.
  - **Example:** If X is the number of heads observed in four flips of a fair coin, the sample space of X is $\{0, 1, 2, 3, 4\}$. The probabilities are given by the binomial distribution. The distribution is $\{0 \to 1/16, 1 \to 1/4, 2 \to 3/8, 3 \to 1/4, 4 \to 1/16\}$.
  - **Example:** Suppose we select a point on the surface of the Earth at random and measure the temperature at that point with an infinitely precise thermometer. The temperature will certainly fall between $-100°C$ and $100°C$, but there are infinitely many values in that range. Thus we can not represent the distribution using a list $\{x \to y, \ldots\}$, as above. Solutions to this problem will be discussed below.
- A random variable is either *qualitative* or *quantitative* depending on the type of value in the sample space. Quantitative random variables express values like temperature, mass, and velocity. Qualitative random variables express values like gender and race.

- The *cumulative distribution function* (CDF) is a way to represent a quantitative distribution. For a random variable X, the CDF is a function $F(t)$ such that $F(t) = P(X \leq t)$. That is, the CDF is a function of t that specifies the probability of observing a value no larger than t.

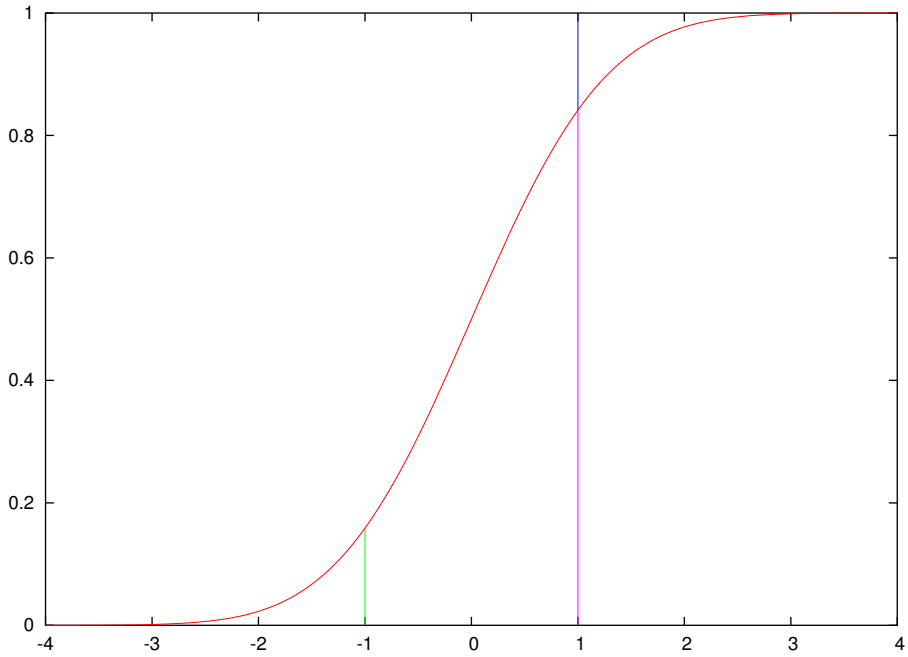
  - **Example:** Suppose X follows a standard normal distribution. You may recall that this distribution has median 0, so that the $P(X \leq 0) = 1/2$ and $P(X \geq 0) = 1/2$. Thus for the standard normal distribution, $F(0) = 1/2$. There is no simple formula for $F(t)$ when $t \neq 0$, but a table of values for $F(t)$ is found in the back of almost any statistics textbook. A plot of $F(t)$ is shown below.



The standard normal CDF

- Any CDF $F(t)$ has the following properties: (i) $0 \leq F(t) \leq 1$, (ii) $F(-\infty) = 0$, (iii) $F(\infty) = 1$, (iv) F is non-decreasing.

- We can read probabilities of the form $P(X \leq t)$ directly from the graph of the CDF. Since $P(X > t) = 1 - P(X \leq t) = 1 - F(t)$, we can also read off a probability of the form $P(X > t)$ directly from a graph of the CDF.

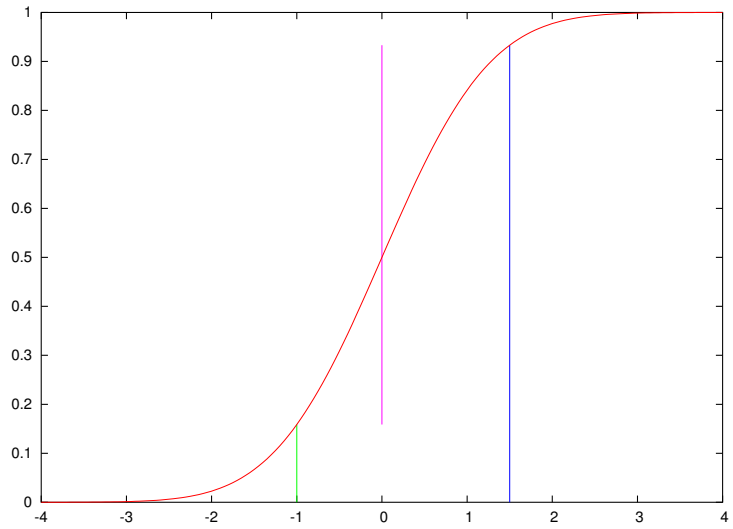


The length of the green line is the probability of observing a value less than -1. The length of the blue line is the probability of observing a value greater than 1. The length of the purple line is the probability of observing a values less than 1.

- If $a \leq b$, for any random variable X

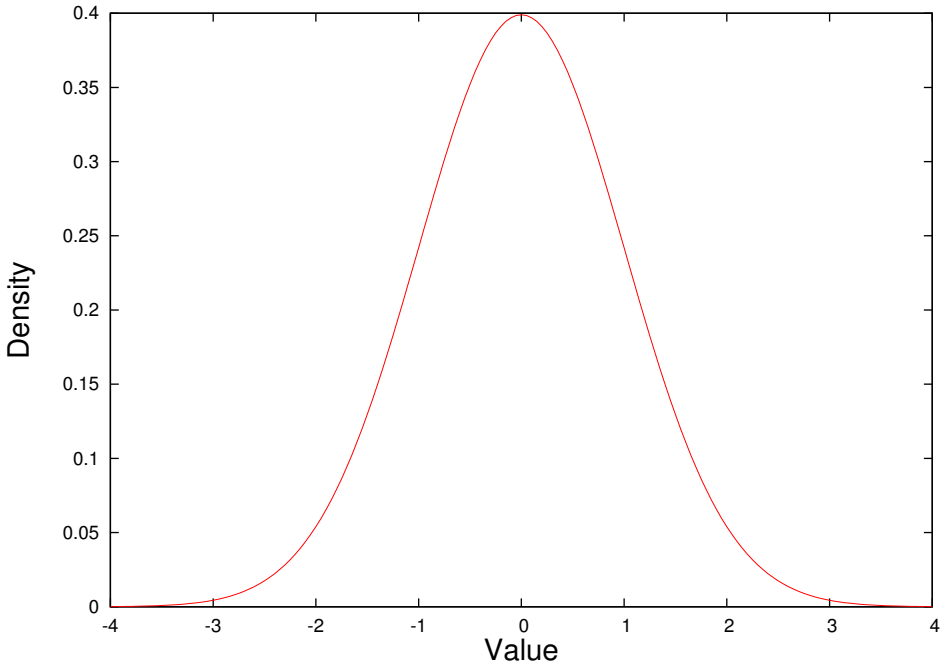
$$P(a < X \leq b) = P(X \leq b) - P(X \leq a) = F(b) - F(a).$$

Thus we can easily determine the probability of observing a value in an interval $(a, b]$ from the CDF.



The length of the purple line is the probability of observing a value between -1 and 1.5.

- If a and b fall in an area where F is very steep, $F(b) - F(a)$ will be relatively large. Thus we are more likely to observe values where F is steep than where F is flat.
- A *probability density function (PDF)* is a different way to represent a probability distribution. The PDF for X is a function $f(x)$ such that the probability of observing a value of X between a and b is equal to the area under the graph of $f(x)$ between a and b. A plot of $f(x)$ for the standard normal distribution is shown below. We are more likely to observed values where f is large than where f is small.



The standard normal PDF

Probability: samples and populations

- If we can repeatedly and independently observe a random variable n times, we have an *independent and identically distributed sample of size n*, or an *iid* sample of size n. This is also called a *simple random sample*, or an *SRS*. (Note that the word *sample* is being used somewhat differently in this context compared to its use in the term *sample space*).

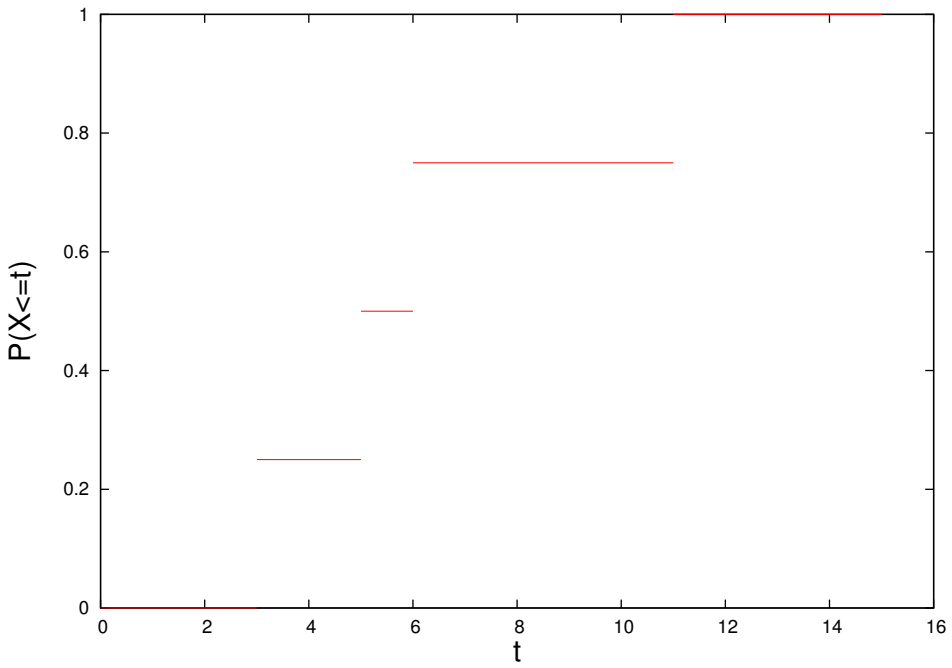
- A central problem in statistics is to answer questions about an unknown distribution called the population based on a simple random sample that was generated by the distribution. This process is called *inference*.

  Specifically, given a numerical characteristic of a distribution, we may wish to estimate the value of that characteristic based on data.

- In an iid sample, each point in the sample space will be observed with a certain *frequency*. For example, if we flip a fair coin 20 times we might observe 13 heads, so the frequency of heads is 13/20. Due to random variation, this frequency differs somewhat from the underlying probability, which is 1/2. If the sample is sufficiently large, frequencies and probabilities will be very similar (this is known as the *law of large numbers*).

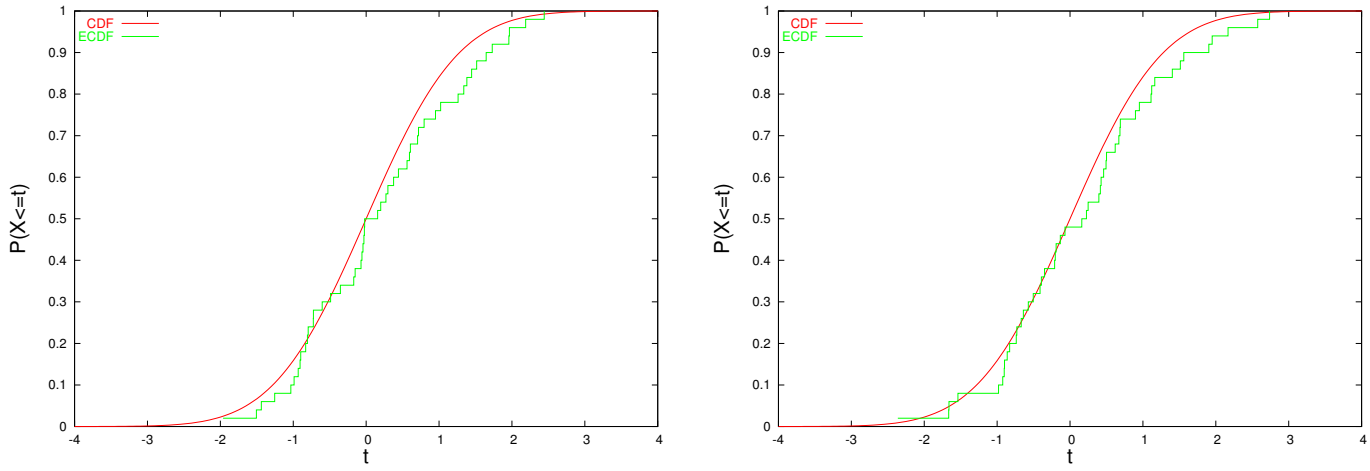
- Since probabilities can be estimated as frequencies, and the CDF is defined in terms of probabilities (i.e. $F(t) = P(X \leq t)$), we can estimate the CDF as the *empirical CDF* (ECDF). Suppose that $X_1, X_2, \ldots, X_n$ are an iid sample. Then the ECDF (evaluated at t) is defined to be the proportion of the X_i that are not larger than t. The ECDF is notated as $\hat{F}(t)$ (in general the symbol $\hat{\star}$ represents an estimate based on an iid sample of a characteristic of the population named $\star$).

  - **Example:** Suppose we observe a sample of size $n = 4$ whose sorted values are $3, 5, 6, 11$. Then $\hat{F}(t)$ is equal to: 0 for $t < 3$, 1/4 for $3 \leq t < 5$, 1/2 for $5 \leq t < 6$, 3/4 for $6 \leq t < 11$, and 1 for $t \geq 11$.



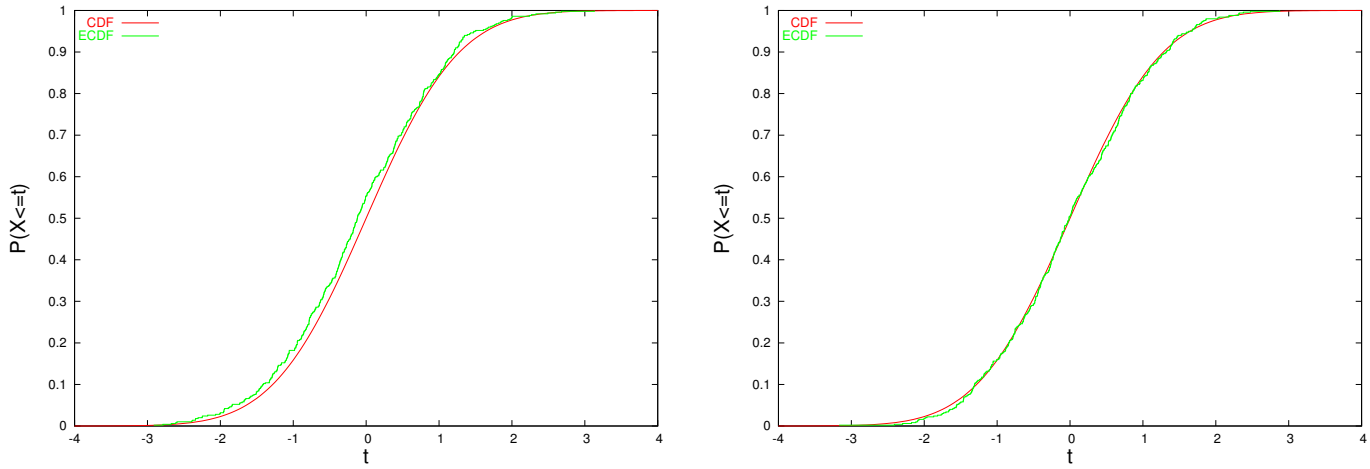
The ECDF for the data set {3, 5, 6, 11}

- Since the ECDF is a function of the sample, which is random, if we construct two ECDF's for two samples from the same distribution, the results will differ (even through the CDF's from the underlying population are the same). This is called *sampling variation*. The next figure shows two ECDF's constructed from two independent samples of size 50 from a standard normal population.



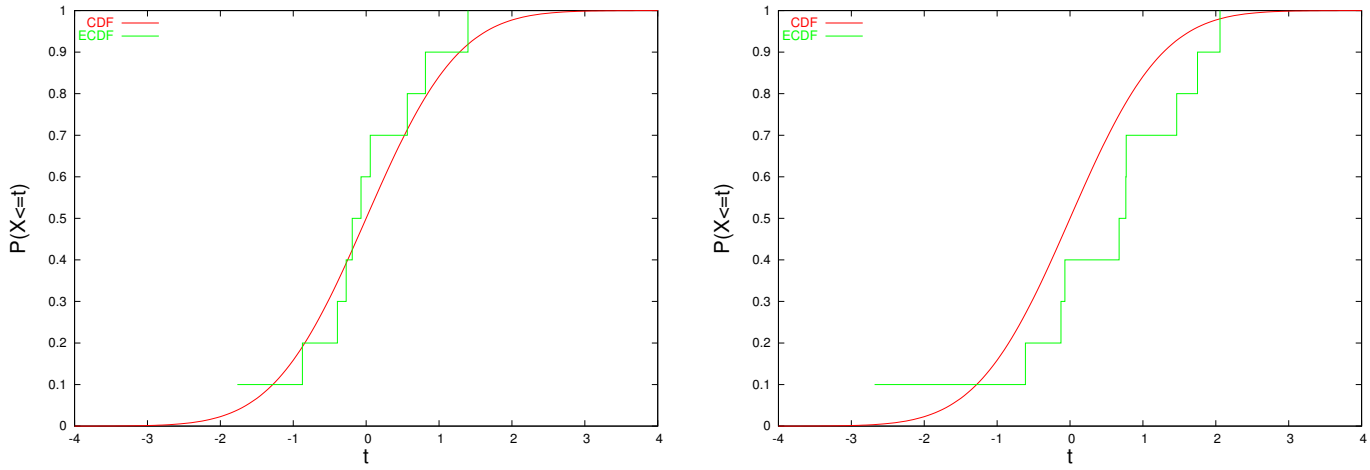
Two ECDF's for standard normal samples of size 50 (the CDF is shown in red)

- The sampling variation gets smaller as the sample size increases. The following figure shows ECDF's based on SRS's of size $n = 500$.



Two ECDF's for standard normal samples of size 500 (the CDF is shown in red)

- The sampling variation gets larger as the sample size decreases. The following figure shows ECDF's based on SRS's of size $n = 10$.



Two ECDF's for standard normal samples of size 10 (the CDF is shown in red)

- Given a SRS $X_1, \ldots, X_n$, a histogram formed from the SRS is an estimate of the PDF. To construct a histogram, select a bin width $\Delta > 0$, and let $H(x)$ be the function such that when $(k-1)\Delta \leq x < k\Delta$, $H(x)$ is the number of observed X_i that fall between $(k-1)\Delta$ and $k\Delta$.

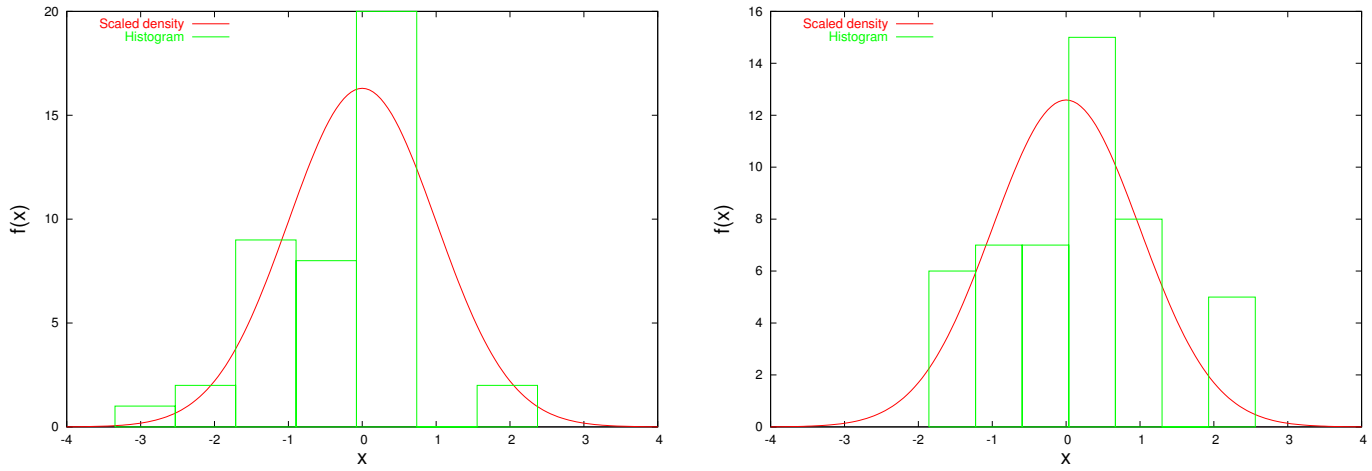
  - To directly compare a density and a histogram they must be put on the same scale. A density is based on a sample of size 1, so to compare it to a histogram based on n observations using bins with width Δ, the density must be scaled by Δn.

- There is no single best way to select Δ. A rule of thumb for the number of bins is

$$\Delta = \frac{R}{\log_2(n) + 1},$$

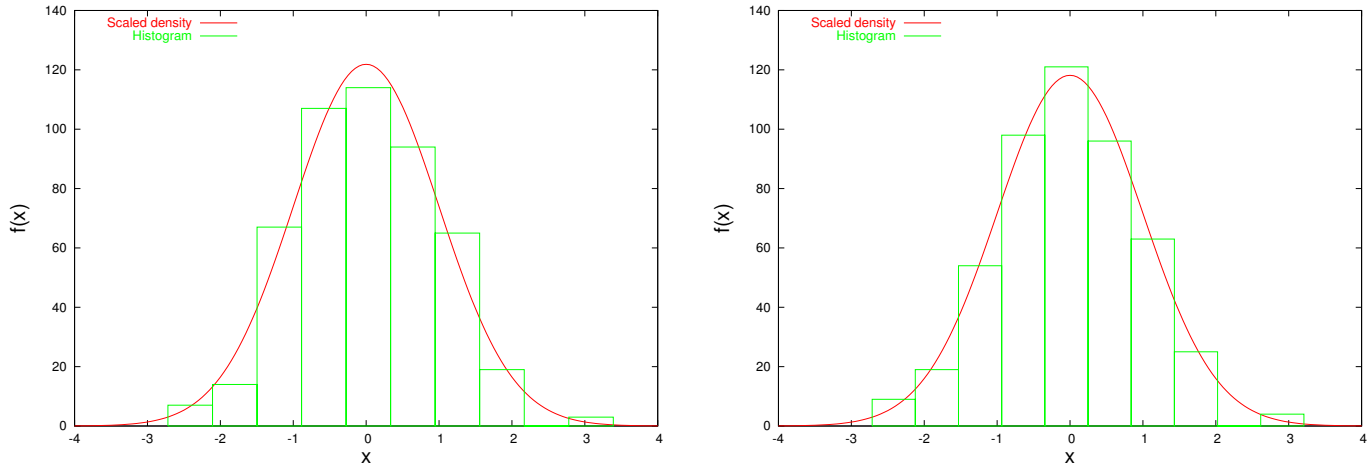
  where n is the number of data points and R is the range of the data (the greatest value minus the least value). This can be used to produce a reasonable value for Δ.

- Just as with the ECDF, sampling variation will cause the histogram to vary if the experiment is repeated. The next figure shows two replicates of a histogram generated from an SRS of 50 standard normal random draws.



Two histograms for standard normal samples of size 50 (the scaled density is shown in red)

- As with the ECDF, larger sample sizes lead to less sampling variation. This is illustrated in comparing the previous figure to the next figure.



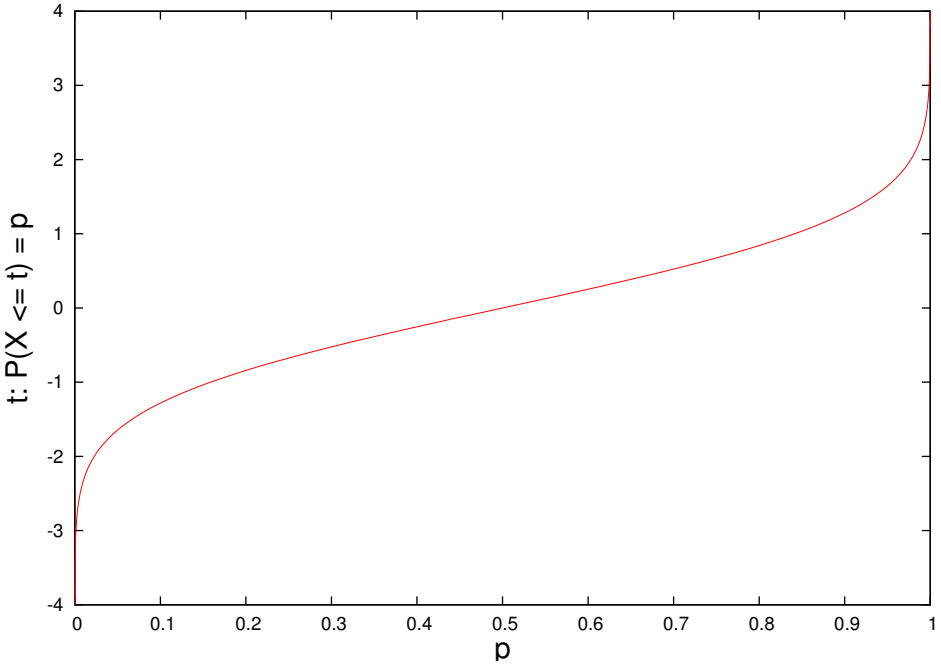
Two histograms for standard normal samples of size 500 (the scaled density is shown in red)

- The *quantile function* is the inverse of the CDF. It is the function $Q(p)$ such that

$$F(Q(p)) = P(X \leq Q(p)) = p,$$

  where $0 \leq p \leq 1$. In words, $Q(p)$ is the point in the sample space such that with probability p the observation will be less than or equal to $Q(p)$. For example, $Q(1/2)$ is the median: $P(X \leq Q(1/2)) = 1/2$, and the 75^{th} percentile is $Q(3/4)$.

- A plot of the quantile function is just a plot of the CDF with the x and y axes swapped. Like the CDF, the quantile function is non-decreasing.

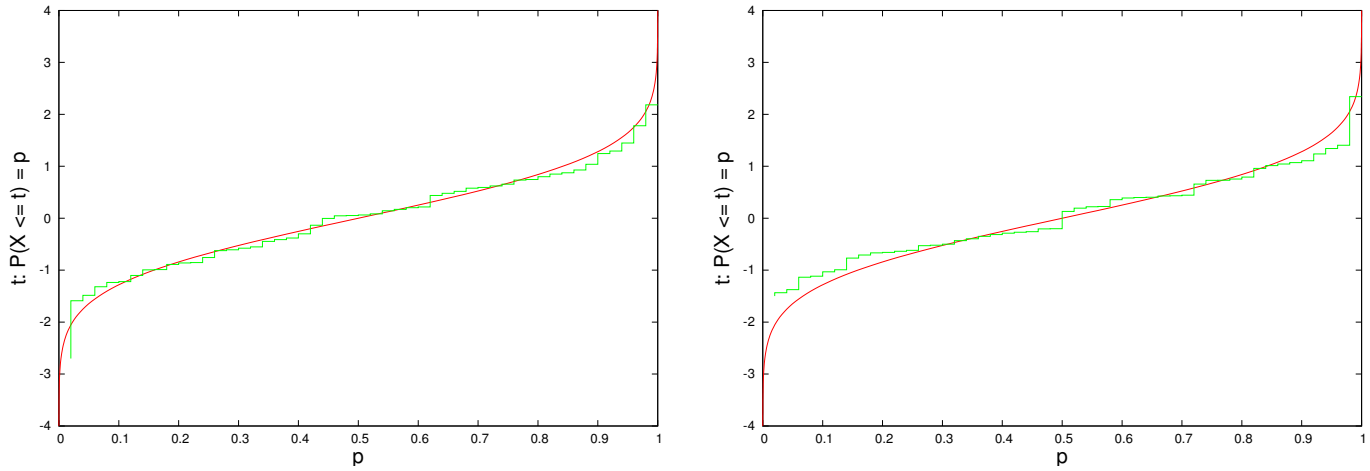


The standard normal quantile function

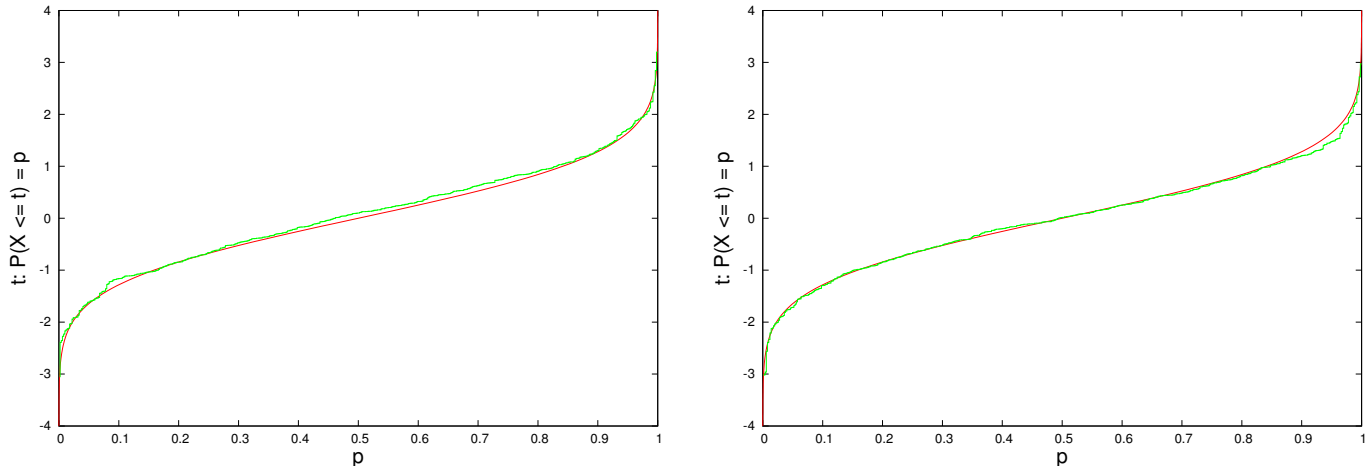
- Suppose we observe an SRS $X_1, X_2, \ldots, X_n$. Sort these values to give $X_{(1)} \leq X_{(2)} \leq \cdots \leq X_{(n)}$ (these are called the *order statistics*). The frequency of observing a value less than or equal to $X_{(k)}$ is k/n. Thus it makes sense to estimate $Q(k/n)$ with $X_{(k)}$, i.e. $\widehat{Q}(k/n) = X_{(k)}$.
- It was easy to estimate $Q(p)$ for $p = 1/n, 2/n, \ldots, 1$. To estimate $Q(p)$ for other values of p, we use *interpolation*. Suppose $k/n < p < (k+1)/n$. Then $\widehat{Q}(p)$ should be between $\widehat{Q}(k/n)$ and $\widehat{Q}((k+1)/n)$ (i.e. between $X_{(k)}$ and $X_{(k+1)}$). To estimate $Q(p)$, we draw a line between the points $(k/n, X_{(k)})$ and $((k+1)/n, X_{(k+1)})$ in the x-y plane. According to the equation for this line, we should estimate $Q(p)$ as:

$$\widehat{Q}(p) = n\left((p - k/n)X_{(k+1)} + ((k+1)/n - p)X_{(k)}\right).$$

  Finally, for the special case $p < 1/n$ set $Q(p) = X_{(1)}$. (There are many slightly different ways to define this interpolation. This is the definition that will be used in this course.)
- The following two figures show empirical quantile functions for standard normal samples of sizes 50 and 500.



Two empirical quantile functions for standard normal samples of size 50 (the population quantile function is shown in red)



Two histograms for standard normal samples of size 500 (the population quantile function is shown in red)

Measures of location

- When summarizing the properties of a distribution, the key features of interest are generally: the most typical value and the level of variability.
- A measure of the most typical value is often called a *measure of location*. The most common measure of location is the *mean*, denoted μ. If $f(x)$ is a density function, then the mean of the distribution is $\mu = \int x f(x) dx$.
- If the distribution has finitely many points in its sample space, it can be notated $\{x_1 \to p_1, \ldots, x_n \to p_n\}$, and the mean is $p_1 x_1 + \cdots + p_n x_n$.
- Think of the mean as the *center of mass* of the distribution. If you had an infinitely long board and marked it in inches from $-\infty$ to ∞, and placed an object with mass p_1 at location X_1, an object with mass p_2 at X_2, and so on, then the mean will be the point at which the board balances.
- The mean as defined above should really be called the population mean, since it is a function of the distribution rather than a sample from the distribution. If we want to estimate the population mean based on a SRS $X_1, \ldots, X_n$, we use the *sample mean*, which is the familiar average: $\bar{X} = (X_1 + \cdots + X_n)/n$. This may also be denoted $\hat{\mu}$.

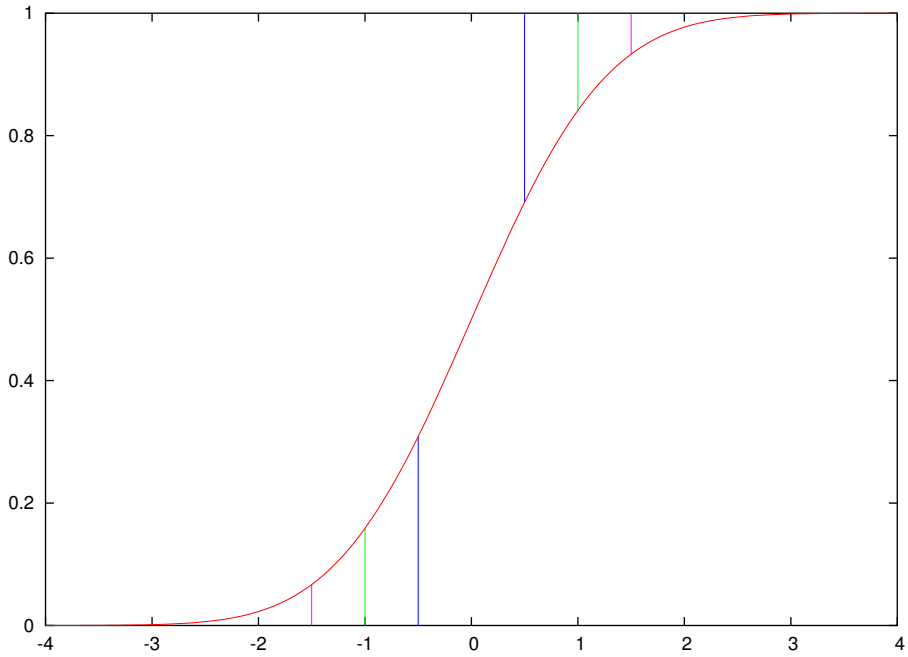
  Note that the population mean is sometimes called the expected value.
- Although the mean is a mathematical function of the CDF and of the PDF, it is not easy to determine the mean just by visually inspecting graphs of these functions.
- An alternative measure of location is the *median*. The median can be easily determined from the quantile function, it is $Q(1/2)$. It can also be determined from the CDF by moving horizontally from $(0, 1/2)$ to the intersection with the CDF, then moving vertically down to the x axis. The x coordinate of the intersection point is the median. The population median can be estimated by the sample median $\widehat{Q}(1/2)$ (defined above).
- Suppose X is a random variable with median θ. Then we will say that X has a symmetric distribution if

$$P(X < \theta - c) = P(X > \theta + c)$$

  for every value of c. An equivalent definition is that

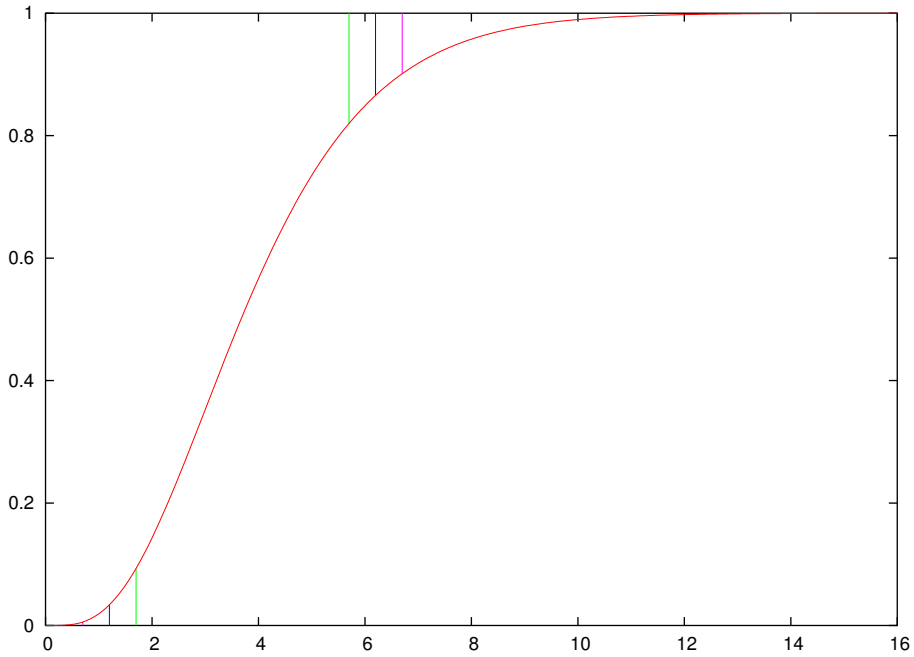
$$F(\theta - c) = 1 - F(\theta + c).$$

  In a symmetric distribution the mean and median are equal. The density of a symmetric distribution is geometrically symmetric about its median. The histogram of a symmetric distribution will be approximately symmetric (due to sampling variation).



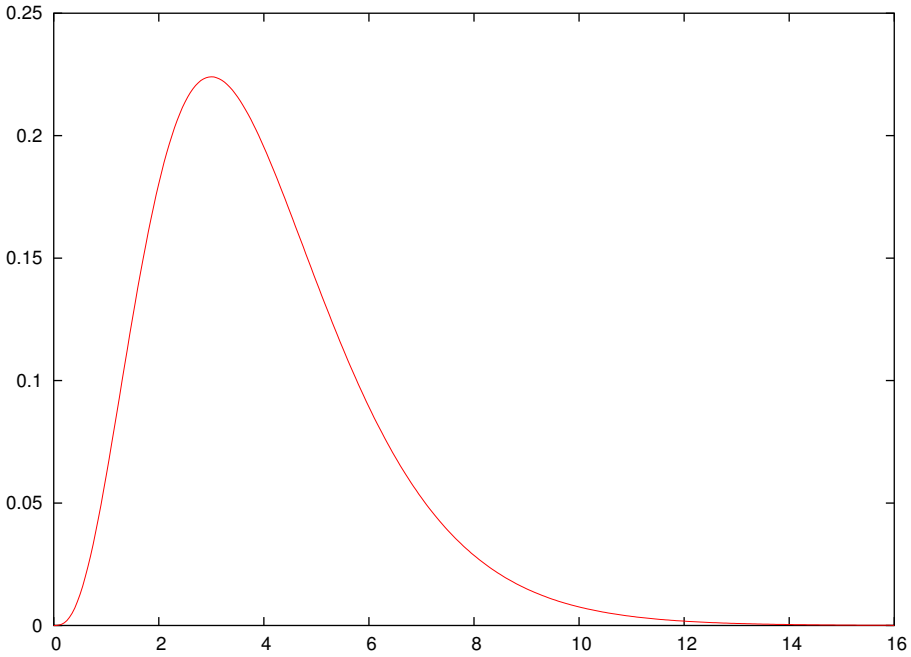
The standard normal CDF. The fact that this CDF corresponds to a symmetric distribution is reflected in the fact that lines of the same color have the same length.

- Suppose that for some values $c > 0$, $P(X > \theta + c)$ is much larger than $P(X < \theta - c)$. That is, we are much more likely to observe values c units larger than the median than values c units smaller than the median. Such a distribution is right-skewed.



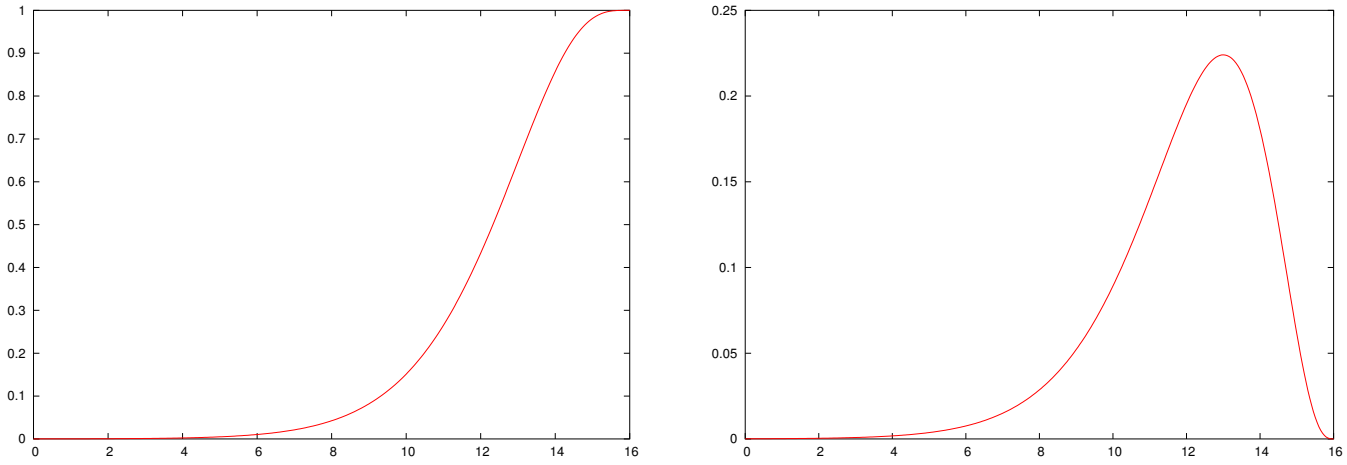
A right-skewed CDF. The fact that the vertical lines on the right are longer than the corresponding vertical lines on the left reflects the fact that the distribution is right-skewed.

The following density function is for the same distribution as the preceeding CDF. Right-skewed distributions are characterized by having long "right tails" in their density functions.



A right-skewed density.

- If $P(X < \theta - c)$ is much larger than $P(X > \theta + c)$ for values of $c > 0$, then the distribution is left-skewed. The following figures show a CDF and density for a left-skewed distribution.



A left-skewed CDF (left) and a left-skewed density (right).

- In a right-skewed distribution, the mean is greater than the median. In a left-skewed distribution, the median is greater than the mean. In a symmetric distribution, the mean and median are equal.

Measures of scale

- A *measure of scale* assesses the level of variability in a distribution. The most common measure of scale is the standard deviation, denoted σ. If $f(x)$ is a density function then $\sigma = \sqrt{\int (x-\mu)^2 f(x) dx}$ is the standard deviation.
- If the distribution has finitely many points in its sample space $\{x_1 \to p_1, \ldots, x_n \to p_n\}$ (notation as used above), then the standard deviation is $\sigma = \sqrt{p_1(x_1-\mu)^2 + \cdots + p_n(x_n-\mu)^2}$.
- The square of the standard deviation is the *variance*, denoted σ^2.
- The standard deviation (SD) measures the distance between a typical observation and the mean. Thus if the SD is large, observations tend to be far from the mean while if the SD is small observations tend to be close to the mean. This is why the SD is said to measure the variability of a distribution.
- If we have data $X_1, \ldots, X_n$ and wish to estimate the population standard deviation, we use the *sample standard deviation*:

$$\hat{\sigma} = \sqrt{\left((X_1 - \bar{X})^2 + \cdots + (X_n - \bar{X})^2\right)/(n-1)}.$$

  It may seem more natural to use n rather than $n-1$ in the denominator. The result is similar unless n is quite small.
- The scale can be assessed visually based on the histogram or ECDF. A relatively wider histogram or a relatively flatter ECDF suggests a more variable distribution. We must say "suggests" because due to the sampling variation in the histogram and ECDF, we can not be sure that what we are seeing is truly a property of the population.
- Suppose that X and Y are two random variables. We can form a new random variable $Z = X + Y$. The mean of Z is the mean of X plus the mean of Y: $\mu_Z = \mu_X + \mu_Y$. If X and Y are independent (to be defined later), then the variance of Z is the variance of X plus the variance of Y: $\sigma_Z^2 = \sigma_X^2 + \sigma_Y^2$.

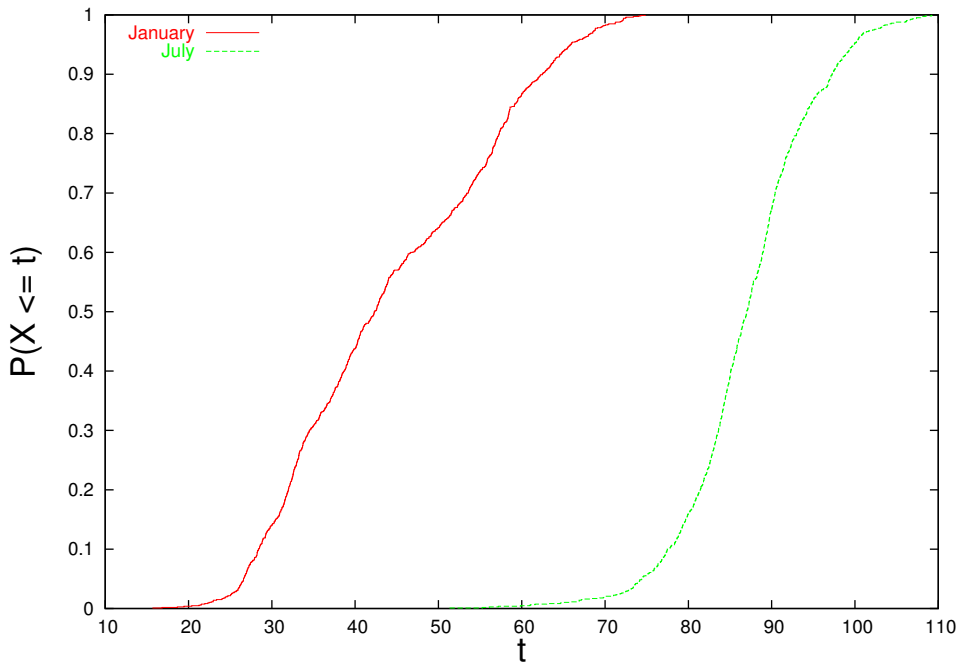
Resistance

- Suppose we observe data $X_1, \ldots, X_{100}$, so the median is $X_{(50)}$ (recall the definition of order statistic given above). Then suppose we observe one additional value Z and recompute the median based on $X_1, \ldots, X_{100}, Z$. There are three possibilities: (i) $Z < X_{(50)}$ and the new median is $(X_{(49)} + X_{(50)})/2$, (ii) $X_{(50)} \leq Z \leq X_{(51)}$, and the new median is $(X_{(50)} + Z)/2$, or (iii) $Z > X_{(51)}$ and the new median is $(X_{(50)} + X_{(51)})/2$. In any case, the new median must fall between $X_{(49)}$ and $X_{(51)}$. When a new observation can only change the value of a statistic by a finite amount, the statistic is said to be *resistant*.
- On the other hand, the mean of $X_1, \ldots, X_{100}$ is $\bar{X} = (X_1 + \cdots + X_{100})/100$, and if we observe one additional value Z then the mean of the new data set is $100\bar{X}/101 + Z/101$. Therefore depending on the value of Z, the new mean can be any number. Thus the sample mean is not resistant.
- The standard deviation is not resistant. A resistant estimate of scale is the *interquartile range* (IQR), which is defined to be $Q(3/4) - Q(1/4)$. It is estimated by the sample IQR, $\hat{Q}(3/4) - \hat{Q}(1/4)$.

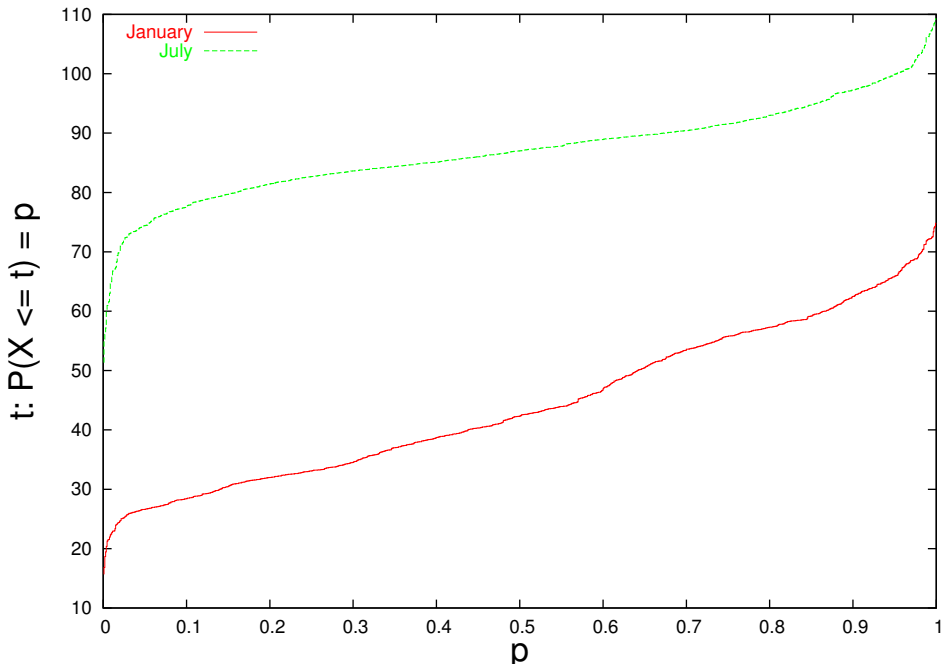
Comparing two distributions graphically

- One way to graphically compare two distributions is to plot their CDF's on a common set of axes. Two key features to look for are
  - The right/left position of the CDF (positions further to the right indicate greater location values).
  - The steepness (slope) of the CDF. A steep CDF (one that moves from 0 to 1 very quickly) suggests a less variable distribution compared to a CDF that moves from 0 to 1 more gradually.
- Location and scale characteristics can also be seen in the quantile function.
  - The vertical position of the quantile function (higher positions indicate greater location values).
  - The steepness (slope) of the quantile function. A steep quantile function suggests a more variable distribution compared to a quantile function that is less steep.
- The following four figures show ECDF's and empirical quantile functions for the average daily maximum temperature over certain months in 2002. Note that January is (of course) much colder than July, and (less obviously) January is more variable than July. Also, the distributions in April and November are very similar (April is a bit colder).

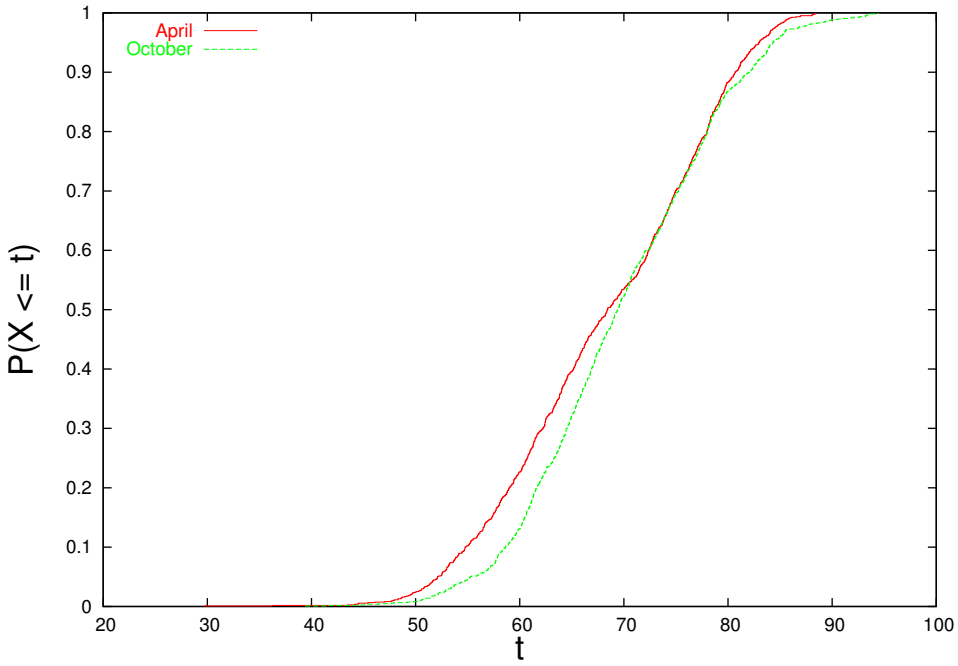
  Can you explain why January is more variable than July?



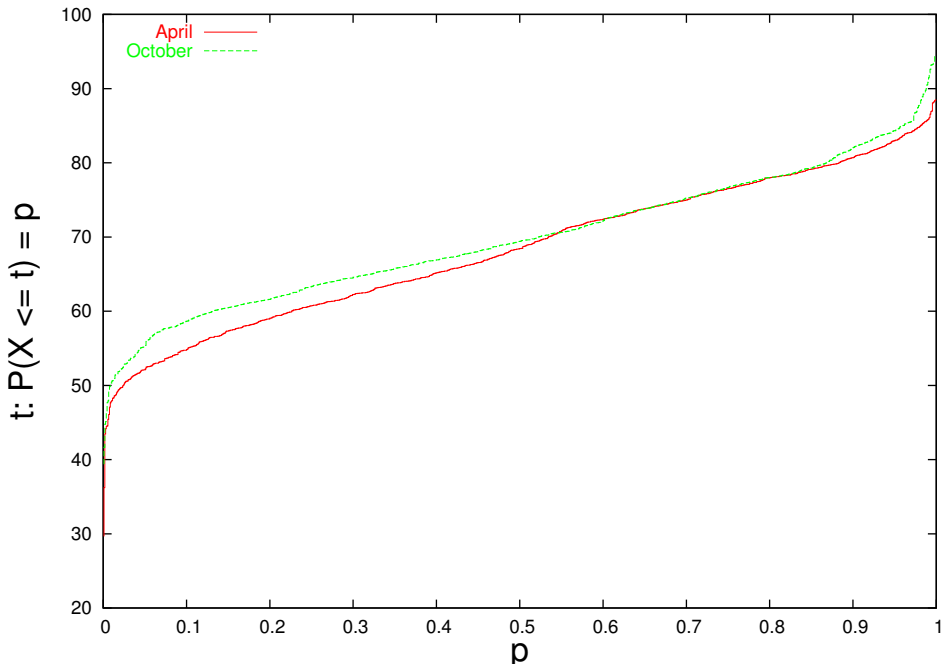
The CDF's for January and July (average daily maximum temperature).



The quantile functions for January and July (average daily maximum temperature).

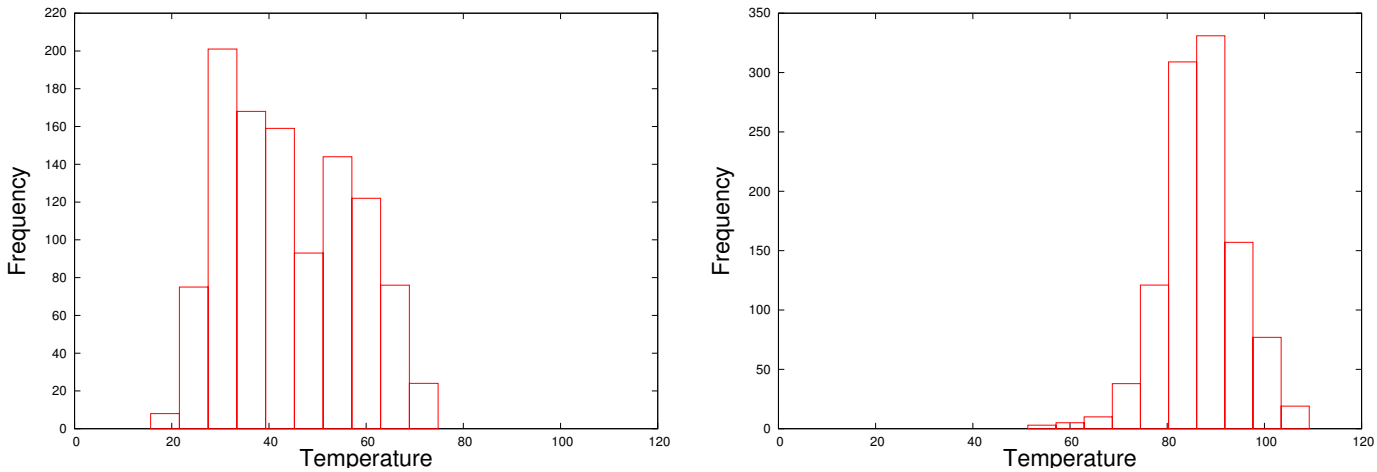


The CDF's for April and October (average daily maximum temperature).

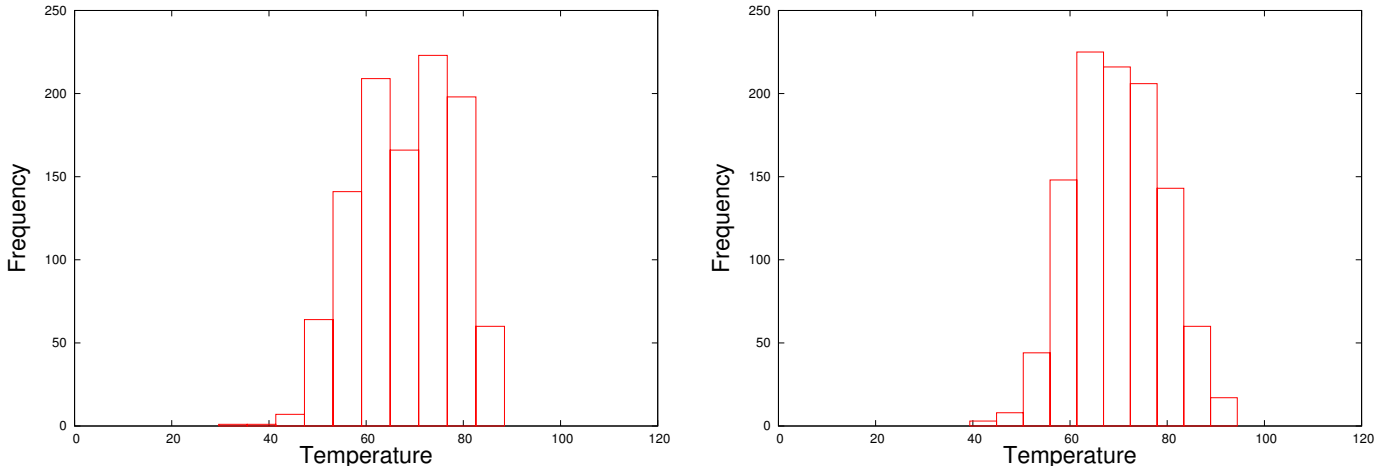


The quantile functions for April and October (average daily maximum temperature).

- Comparisons of two distributions can also be made using histograms. Since the histograms must be plotted on separate axes, the comparisons are not as visually clear.

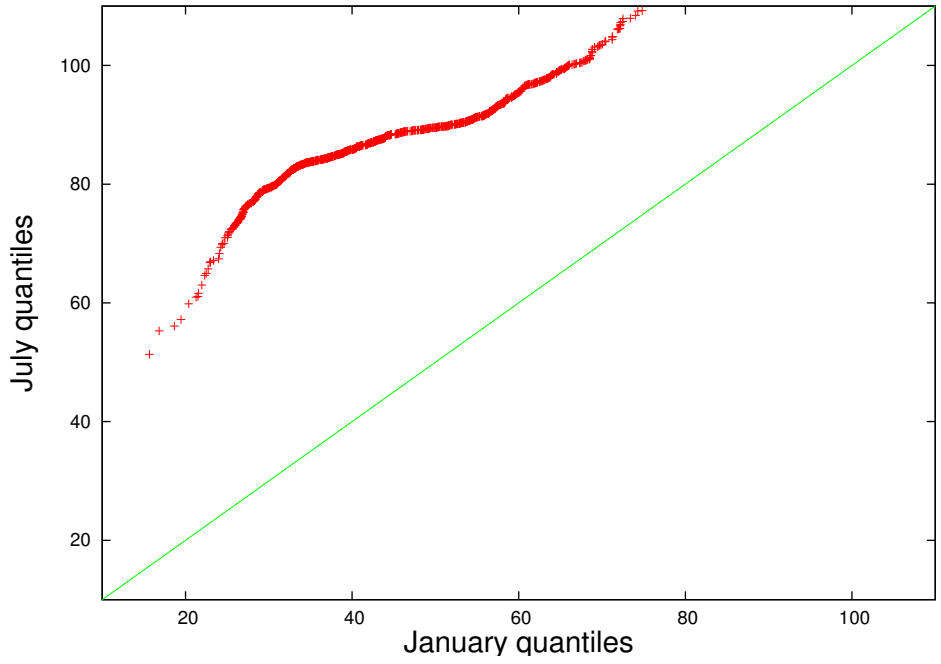


Histograms for January and July (average daily maximum temperature).



Histograms for April and October (average daily maximum temperature).

- The standard graphical method for comparing two distributions is a *quantile-quantile* (QQ) plot. Suppose that $\hat{Q}_X(p)$ is the empirical quantile function for $X_1, \ldots, X_m$ and $\hat{Q}_Y(p)$ is the empirical quantile function for $Y_1, \ldots, Y_n$. If we make a scatterplot of the points $(\hat{Q}_X(p), \hat{Q}_Y(p))$ in the plane for every $0 < p < 1$ we get something that looks like the following:



QQ plot of average daily maximum temperature (July vs. January).

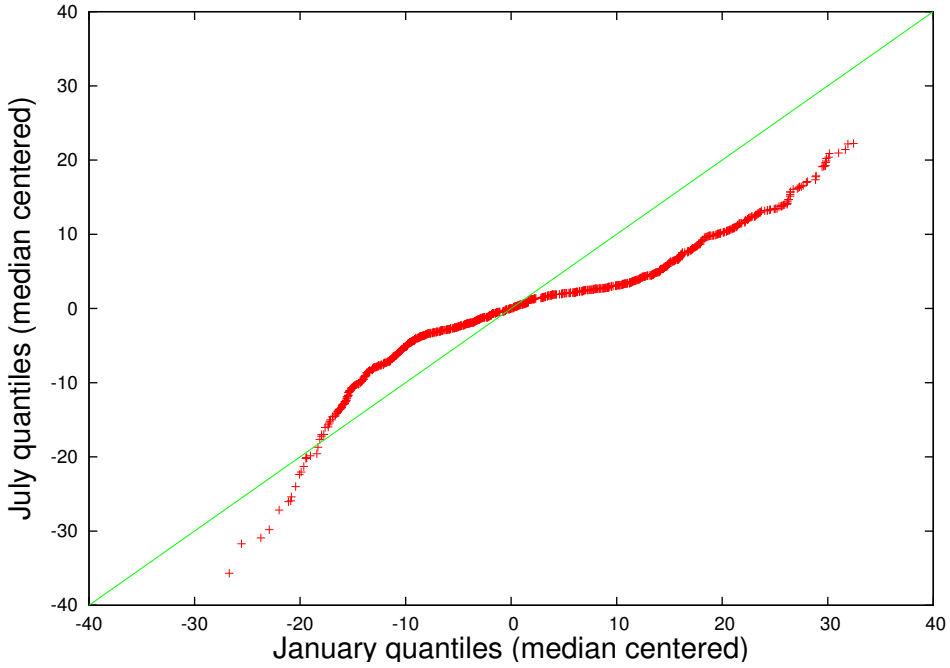
- The key feature in the plot is that every quantile in July is greater than the corresponding quantile in January.

  More subtly, since the slope of the points is generally shallower than 45°, we infer that January temperatures are more variable than July temperatures (if the slope were much greater than 45° then we would infer that July temperatures are more variable than January temperatures).

- If we take it as obvious that it is warmer in July than January, we may wish to modify the QQ plot to make it easier to make other comparisons.

  We may *median center* the data (subtract the median January temperature from every January temperature and similarly with the July temperatures) to remove location differences.

  In the median centered QQ plot, it is very clear that January temperatures are more variable throughout most of the range, although at the low end of the scale there are some points that do not follow this trend.



QQ plot of median centered average daily maximum temperature (July vs. January).

- A QQ plot can be used to compare the empirical quantiles of a sample $X_1, \ldots, X_n$ to the quantiles of a distribution such as the standard normal distribution. Such a plot is called a normal probability plot.

  The main application of a normal probability plot is to assess whether the tails of the data are thicker, thinner, or comparable to the tails of a normal distribution.

  The tail thickness determines how likely we are to observe extreme values. A thick right tail indicates an increased likelihood of observing extremely large values (relative to a normal distribution). A thin right tail indicates a decreased likelihood of observing extremely large values.

  The left tail has the same interpretation, but replace "extremely large" with "extremely small" (where "extremely small" means "far in the direction of $-\infty$").
- To assess tail thickness/thinness from a normal probability plot, it is important to note whether the data quantiles are on the X or Y axis. Assuming that the data quantiles are on the Y axis:

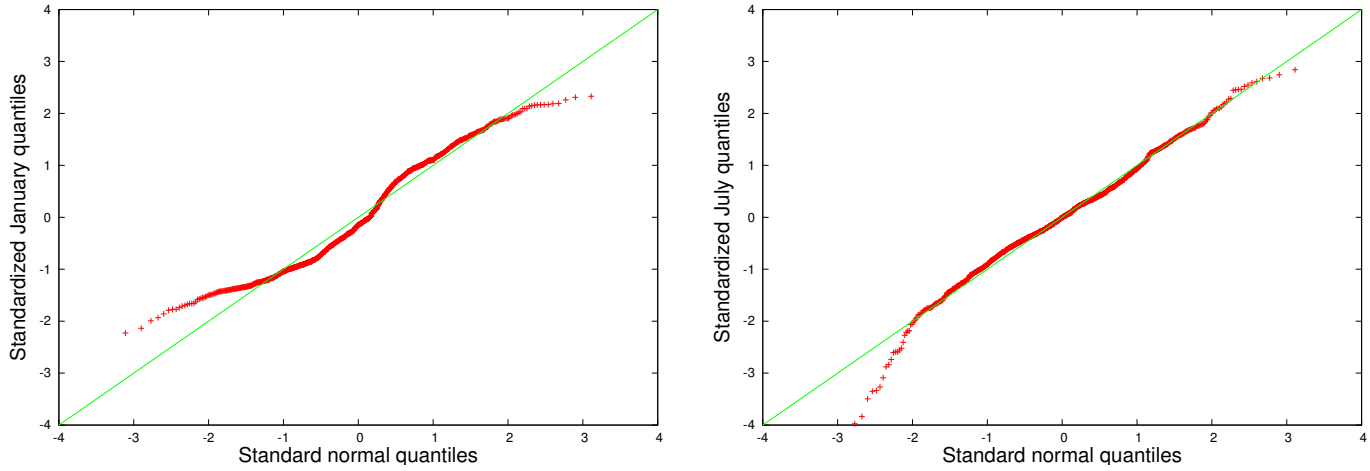
  - A thick right tail falls above the 45° diagonal, a thin right tail falls below the 45° diagonal.
  - A thick left tail falls below the 45° diagonal, a thin left tail falls above the 45° diagonal.

  If the data quantiles are on the X axis, the opposite holds (thick right tails fall below the 45°, etc.).
- Suppose we would like to assess whether the January or July maximum temperatures are normally distributed. To accomplish this, perform the following steps.

  - First we standardize the temperature data, meaning that for each of the two months, we compute the sample mean $\hat{\mu}$ and the sample standard deviation $\hat{\sigma}$, then transform each value using

    $$Z \to (Z - \hat{\mu})/\hat{\sigma}.$$

    Once this has been done, then the transformed values for each month will have sample mean 0 and sample standard deviation 1, and hence can be compared to a standard normal distribution.
  - Next we construct a plot of the temperature quantiles (for standardized data) against the corresponding population quantiles of the standard normal distribution. The simplest way to proceed is to plot $Z_{(k)}$ (where $Z_1, Z_2, \ldots$ are the standardized temperature data) against $Q(k/n)$, where Q is the standard normal quantile function.



QQ plot of standardized average daily maximum temperature in January (left) and July (right) against standard normal quantiles.

- In both cases, the tails for the data are roughly comparable to normal tails. For January both tails are slightly thinner than normal, and the left tail for July is slightly thicker than normal.

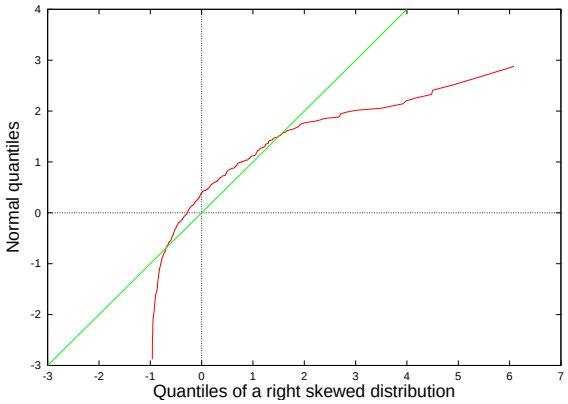
  The atypical points for July turn out to correspond to a few stations at very high elevations that are unusually cold in summer, e.g. Mount Washington and a few stations in the Rockies.
- Normal probability plots can also be used to detect skew.

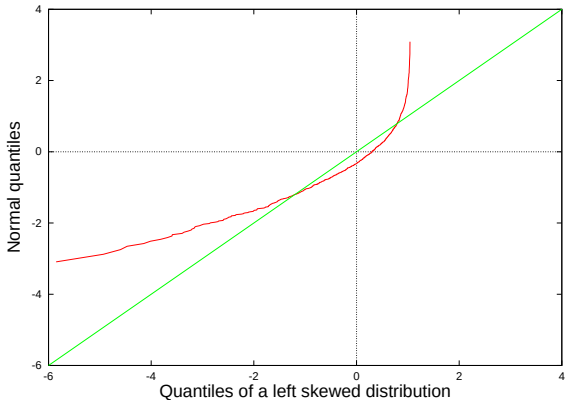
  The following two figures show the general pattern for the normal probability plot for left skewed and for right skewed distributions.

  The key to understanding these figures is to consider the extreme (largest and smallest) quantiles.

  - In a right skewed distribution, the largest quantiles will be much larger compared to the corresponding normal quantiles.
  - In a left skewed distribution, the smallest quantiles will be much smaller compared to the corresponding normal quantiles.

  Be sure to remember that "small" means "closer to $-\infty$", not "closer to 0".

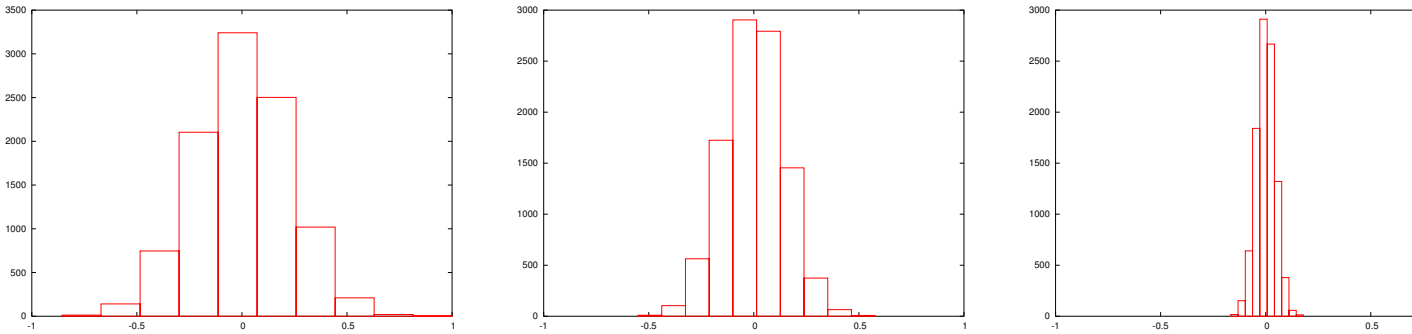




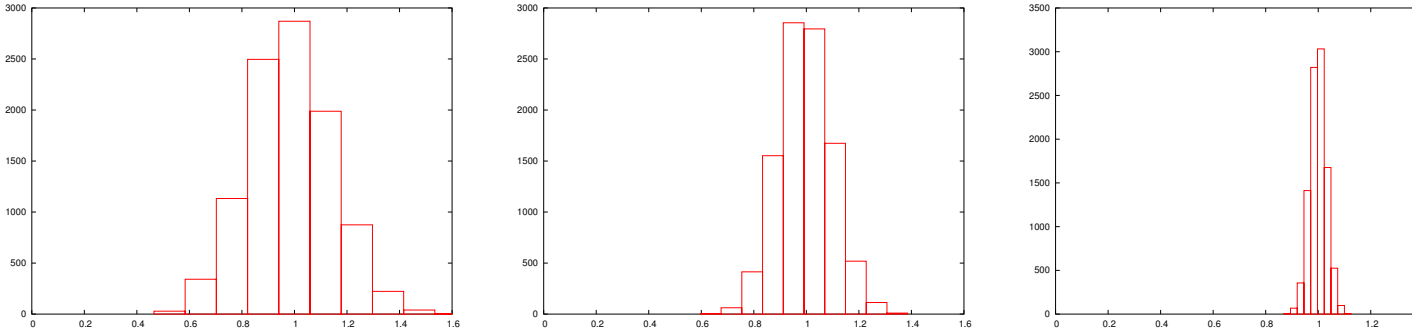
- Note that the data quantiles are on the X axis (the reverse of the preceeding normal probability plots). It is important that you be able to read these plots both ways.

Sampling distributions of statistics

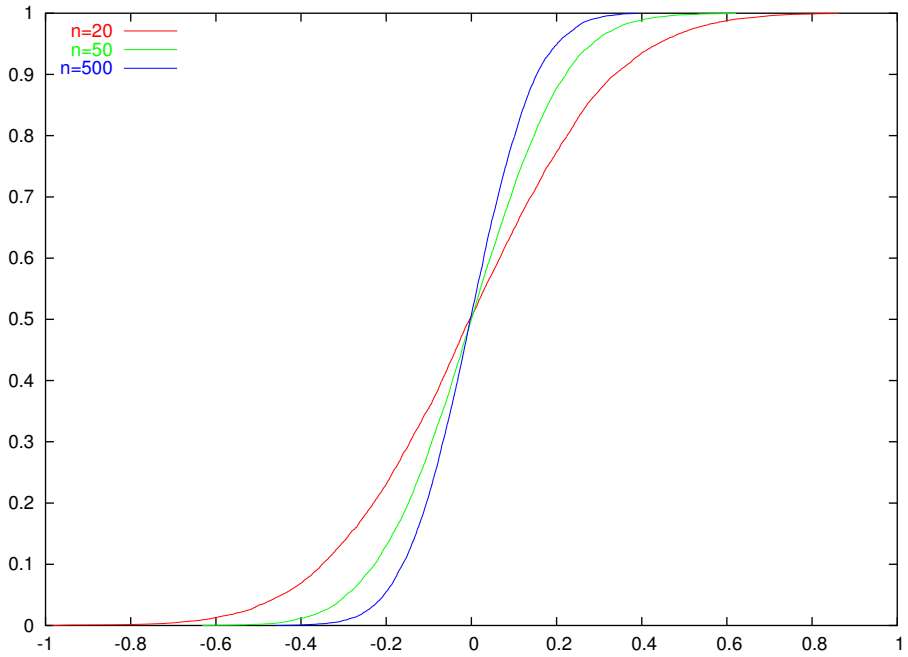
- A statistic is any function of a random variable (i.e. a function of data). For example, the sample mean, sample median, sample standard deviation, and sample IQR are all statistics.
- Since a statistic is formed from data, which is random, a statistic itself is random. Hence a statistic is a random variable, and it has a distribution. The variation in this distribution is referred to as sampling variation.
- The distribution of a statistic is determined by the distribution of the data used to form the statistic. However there is no simple procedure that can be used to determine the distribution of a statistic from the distribution of the data.
- Suppose that $\bar{X}$ is the average of a SRS $X_1, \ldots, X_n$. The mean and standard deviation of $\bar{X}$ are related to the mean μ and standard deviation σ of X_i as follows. The mean of $\bar{X}$ is μ and the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$.
- Many simple statistics are formed from a SRS, for example the sample mean, median, standard deviation, and IQR. For such statistics, the key characteristic is that the sampling variation becomes smaller as the sample size increases. The following figures show examples of this phenomenon.



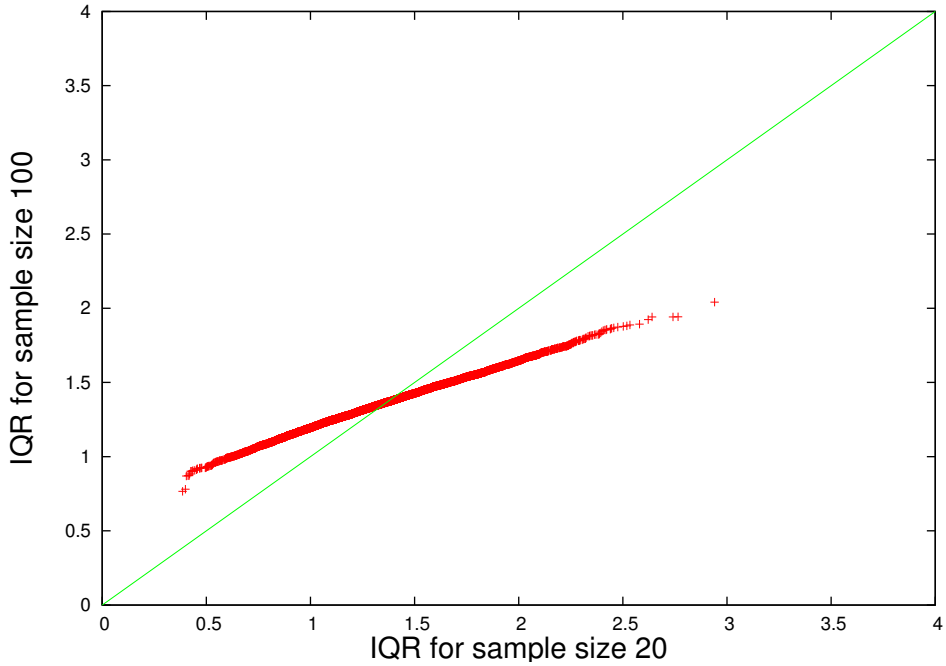
Sampling variation of the sample mean for standard normal SRS's of size 20, 50, and 500.



Sampling variation of the sample standard deviation for standard normal SRS's of size 20, 50, and 500.



ECDF's showing the sampling variation in the sample median for standard normal SRS's of size 20, 50, and 500.



QQ plot the showing the sampling variation in the sample IQR for standard normal SRS's of size 20 (x axis) and 100 (y axis). The true value is 1.349.

- In the case of the sample mean, we can directly state how the variation decreases as a function of the sample size: for an SRS of size n, the standard deviation of $\bar{X}$ is $\sigma/\sqrt{n}$, where σ is the standard deviation of one observation.

  The sample size must increase by a factor of 4 to cut the standard deviation in half. Doubling the sample size only reduces $\hat{\sigma}$ by around 30%.

  For other statistics such as the sample median or sample standard deviation, the variation declines with sample size. But it is not easy to give a formula for the standard deviation in terms of sample size.

  For most statistics, it is approximately true that increasing the sample size by a factor of F scales the sample standard deviation by a factor of $1/\sqrt{F}$.

Hypothesis testing

- In most practical data analysis it is possible to carry out inferences (from sample to population) based on graphical techniques (e.g. using the empirical CDF and quantile functions and the histogram).

  This type of inference may be considered informal, since it doesn't involve making quantitative statements about the likelihood that certain characteristics of the population hold.
- In certain cases it is important to make quantitative statements about the degree of uncertainty in an inference. This requires a formal and quantitative approach to inference.
- In the standard setup we are considering *hypotheses*, which are statements about a population. For example, the statement that the mean of a population is positive is a hypothesis.

  More concretely, we may be comparing incomes of workers with a BA degree to incomes of workers with an MA degree, and our hypothesis may be that the mean MA income minus the mean BA income is positive.

  Note that hypotheses are always statements about populations, not samples, so the means above are population means.
- Generally we are comparing two hypotheses, which are conventionally referred to as the null hypothesis and the alternative hypothesis.

  If the data are inconclusive or strongly support the null hypothesis, then we decide in favor of the null hypothesis. Only if the data strongly favor the alternative hypothesis do we decide in favor of the alternative hypothesis over the null.

– *Example:* If hypothesis A represents a "conventional wisdom" that somebody is trying to overturn by proposing hypothesis B, then A should be the null hypothesis and B should be the alternative hypothesis. Thus, if somebody is claiming that cigarette smoking is not associated with lung cancer, the null hypothesis would be that cigarette smoking *is* associated with lung cancer, and the alternative would be that it is not. Then once the data are collected and analyzed, if the results are inconclusive, we would stick with the standard view that smoking and lung cancer are related.

  Note that the "conventional wisdom" may change over time. One-hundred years ago smoking was not widely regarded as dangerous, so the null and alternative may well have been switched back then.

– *Example:* If the consequences of mistakenly accepting hypothesis A are more severe than the consequences of mistakenly accepting hypothesis B, then B should be the null hypothesis and A should be the alternative. For example, suppose that somebody is proposing that a certain drug prevents baldness, but it is suspected that the drug may be very toxic. If we adopt the use of the drug and it turns out to be toxic, people may die. On the other hand if we do not adopt the use of the drug and it turns out to be effective and non-toxic, some people will needlessly become bald. The consequence of the first error is far more severe than the consequence of the second error. Therefore we take as the null hypothesis that the drug is toxic, and as the alternative we take the hypothesis that the drug is non-toxic and effective.

  Note that if the drug were intended to treat late stage cancer, the designation would not be as clear because the risks of not treating the disease are as severe as the risk of a toxic reaction (both are likely to be fatal).

– *Example:* If hypothesis A is a much simpler explanation for a phenomenon than hypothesis B, we should take hypothesis A as the null hypothesis and hypothesis B as the alternative hypothesis. This is called the *principle of parsimony*, or *Occam's razor*. Stated another way, if we have no reason to favor one hypothesis over another, the simplest explanation is preferred.

  Note that there is no general theoretical justification for this principal, and it does sometimes happen that the simplest possible explanation turns out to be incorrect.

- Next we need to consider the level of *evidence* in the data for each of the two hypotheses. The standard method is to use a *test statistic* $T(X_1, \ldots, X_n)$ such that extreme values of T indicate evidence for the alternative hypothesis, and non-extreme values of T indicate evidence for the null hypothesis.

  "Extreme" may mean "*closer to* $+\infty$" (a right-tailed test), or "*closer to* $-\infty$" (a left-tailed test), or "*closer to one of* $\pm\infty$", depending on the context. The first two cases are called one-sided tests, while the final case is called a two-sided test.

  The particular definition of "extreme" for a given problem is called the rejection region.

- *Example:* Suppose we are investigating a coin, and the null hypothesis is that the coin is fair (equally likely to land heads or tails) while the alternative is that the coin is unfairly biased in favor of heads. If we observe data $X_1, \ldots, X_n$ where each X_i is H or T, then the test statistic $T(X_1, \ldots, X_n)$ may be the number of heads, and the rejection region would be "large values of T" (since the maximum value of T is n, we might also say "T close to n").

  On the other hand, if the alternative hypothesis was that the coin is unfairly biased in favor of tails, the rejection region would be "small values of T" (since the minimum value of T is zero, we might also say "T close to zero").

  Finally, if if the alternative hypothesis was that the coin is unfairly biased in any way, the rejection region would be "large or small values of T" (T close to 0 or n).

- *Example:* Suppose we are investigating the effect of eating fast food on body shape. We choose to focus on the body mass index $X = \text{weight}/\text{height}^2$, which we observe for people $X_1, \ldots, X_m$ who never eat fast food and people $Y_1, \ldots, Y_n$ who eat fast food three or more times per week. Our null hypothesis is that the two populations have the same mean BMI, and the alternative hypothesis is that people who eat fast food have a higher mean BMI.

  We shall see that a reasonable test statistic is

  $$T = (\bar{Y} - \bar{X})/\sqrt{\hat{\sigma}_X^2/m + \hat{\sigma}_Y^2/n}$$

  where $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are the sample standard deviations for the X_i and the Y_i respectively). The rejection region will be "large values of T".
- In making a decision in favor of the null or alternative hypothesis, two errors are possible:

  A type I error, or false positive occurs when we decide in favor of the alternative hypothesis when the null hypothesis is true.

  A type II error, or false negative occurs when we decide in favor of the null hypothesis when the alternative hypothesis is true.

  According to the way that the null and alternative hypotheses are designated, a false positive is a more undesirable outcome than a false negative.
- Once we have a test statistic T and a rejection region, we would like to quantify the amount of evidence in favor of the alternative hypothesis.

  The standard method is to compute the probability of observing a value of T "as extreme or more extreme" than the observed value of T, assuming that the null hypothesis is true.

  This number is called the *p-value*. It is the probability of type I error, or the probability of making a false positive decision, if we decide in favor of the alternative based on our data.

  For a right-tailed test, the p-value is $P(T \geq T_{\text{obs}})$, where T_{obs} denotes the test statistic value computed from the observed data, and T denotes a test statistic value generated by the null distribution.

  Equivalently, the right-tailed p-value is $1 - F(T_{\text{obs}})$, where F is the CDF of T under the null hypothesis.
  For a left-tailed test, the p-value is $P(T \leq T_{\text{obs}})$, or equivalently $F(T_{\text{obs}})$.

  For a two sided test we must locate the "most typical value of T" under the null hypothesis and then consider extreme values centered around this point. Suppose that μ_T is the expected value of the test statistic under the null hypothesis. Then the p-value is

  $$P(|T - \mu_T| > |T_{\text{obs}} - \mu_T|)$$

  which can also be written

  $$P(T < \mu_T - |T_{\text{obs}} - \mu_T|) + P(T > \mu_T + |T_{\text{obs}} - \mu_T|).$$

- *Example:* Suppose we observe 28 heads and 12 tails in 40 flips of a coin. Our observed test statistic value is $T_{\text{obs}} = 28$. You may recall that under the null hypothesis ($P(H) = P(T) = 1/2$) the probability of observing exactly k heads out of 40 flips is $\binom{40}{k}/2^{40}$ (where $\binom{n}{k} = n!/(n-k)!k!$). Therefore the probability of observing a test statistic value of 28 or larger under the null hypothesis (i.e. the p-value) is

$$P(T = 28) + P(T = 29) + \cdots + P(T = 40)$$

which equals

$$\binom{40}{28}/2^{40} + \binom{40}{29}/2^{40} + \cdots + \binom{40}{40}/2^{40}.$$

This value can be calculated on a computer. It is approximately .008, indicating that it is very unlikely to observe 28 or more heads in 40 flips of a fair coin. Thus the data suggests that the coin is not fair, and in particular it is biased in favor of heads.

Put another way, if we decide in favor of the alternative hypothesis, there is $< 1\%$ chance that we are committing a type I error.
An alternative approach to calculating this p-value is to use a normal approximation. Under the null distribution, T has mean $n/2$ and standard deviation $\sqrt{n}/2$ (recall the standard deviation formula for the binomial distribution is $\sigma = \sqrt{np(1-p)}$ and substitute $p = 1/2$).

Thus the standardized test statistic is $T^*_{\text{obs}} = 2(T_{\text{obs}} - n/2)/\sqrt{n}$, which is 2.53 in this case. Since T^*_{obs} has mean 0 and standard deviation 1 we may approximate its distribution with a standard normal distribution. Thus the p-value can be approximated as the probability that a standard normal value exceeds 2.53. From a table of the standard normal distribution, this is seen to be approximately .006, which is close to the true value of (approximately) .008 and can be calculated without the use of a computer.

- *Example:* Again suppose we observe 28 heads out of 40 flips, but now we are considering the two-sided test. Under the null hypothesis, the expected value of T is $\mu_T = n/2 = 20$. Therefore the p-value is $P(|T - 20| \geq |T_{\text{obs}} - 20|)$, or $P(|T - 20| \geq 8)$. To compute the p-value exactly using the binomial distribution we calculate the sum

$$P(T = 0) + \cdots + P(T = 12) + P(T = 28) + \cdots + P(T = 40)$$

which is equal to

$$\binom{40}{0}/2^{40} + \cdots + \binom{40}{12}/2^{40} + \binom{40}{28}/2^{40} + \cdots + \binom{40}{40}/2^{40}.$$

To approximate the p-value using the standard normal distribution, standardize the boundary points of the rejection region (12 and 28) just as T_{obs} was standardized above. This yields ± 2.53. From a normal probability table, $P(Z > 2.53) = P(Z < -2.53) \approx 0.006$, so the p-value is approximately 0.012.

Under the normal approximation, the two-sided p-value will always be twice the on-sided p-value. However for the exact p-values this may not be true.

- *Example:* Suppose we observe BMI's $Y_1, \ldots, Y_{30}$ such that the sample mean and standard deviation are $\bar{Y} = 26$ and $\hat{\sigma}_Y = 4$ and another group of BMI's $X_1, \ldots, X_{20}$ with $\bar{X} = 24$ and $\hat{\sigma}_X = 3$. The test statistic (formula given above) has value 2.02. Under the null hypothesis, this statistic approximately has a standard normal distribution. The probability of observing a value greater than 2.02 (for a right-tailed test) is .022. This is the p-value.

Planning an experiment or study

- When conducting a study, it is important to use a sample size that is large enough to provide a good chance reaching the correct conclusion.

  Increasing the sample size always increases the chances of reaching the right conclusion. However every sample costs time and money to collect, so it is desirable to avoid making an unnecessarily large number of observations.
- It is common to use a p-value cutoff of .01 or .05 to indicate "strong evidence" for the alternative hypothesis. Most people feel comfortable concluding in favor of the alternative hypothesis if such a p-value is found.

  Thus in planning, one would like to have a reasonable chance of obtaining such a p-value if the alternative is in fact true.

  On the other hand, consider yourself lucky if you observe a large p-value when the null is true, because you can cut your losses and move on to a new investigation.
- In many cases, the null hypothesis is known exactly but the precise formulation of the alternative is harder to specify.

  For instance, I may suspect that somebody is using a coin that is biased in favor of heads. If p is the probability of the coin landing heads, it is clear that the null hypothesis should be $p = 1/2$.

  However it is not clear what value of p should be specified for the alternative, beyond that p should be greater than $1/2$.

  The alternative value of p may be left unspecified, or we may consider a range of possible values. The difference between a possible alternative value of p and the null value of p is the effect size.
- If the alternative hypothesis is true, it is easier to get a small p-value when the effect size is large, i.e. for a situation in which the alternative hypothesis is "far" from the null hypothesis. This is illustrated by the following examples.

  - Suppose your null hypothesis is that a coin is fair, and the alternative is $p > 1/2$. An effect size of 0.01 is equivalent to an alternative heads probability of 0.51.

    For reasonable sample sizes, data generated from the null and alternative hypotheses look very similar (e.g., under the null the probability of observing 10/20 heads is ≈ 0.17620 while under the alternative the same probability is ≈ 0.17549).
  - Now suppose your null hypothesis is that a coin is fair, the alternative hypothesis is $p > 1/2$, and the effect size is 0.4, meaning that the alternative heads probability is 0.9.

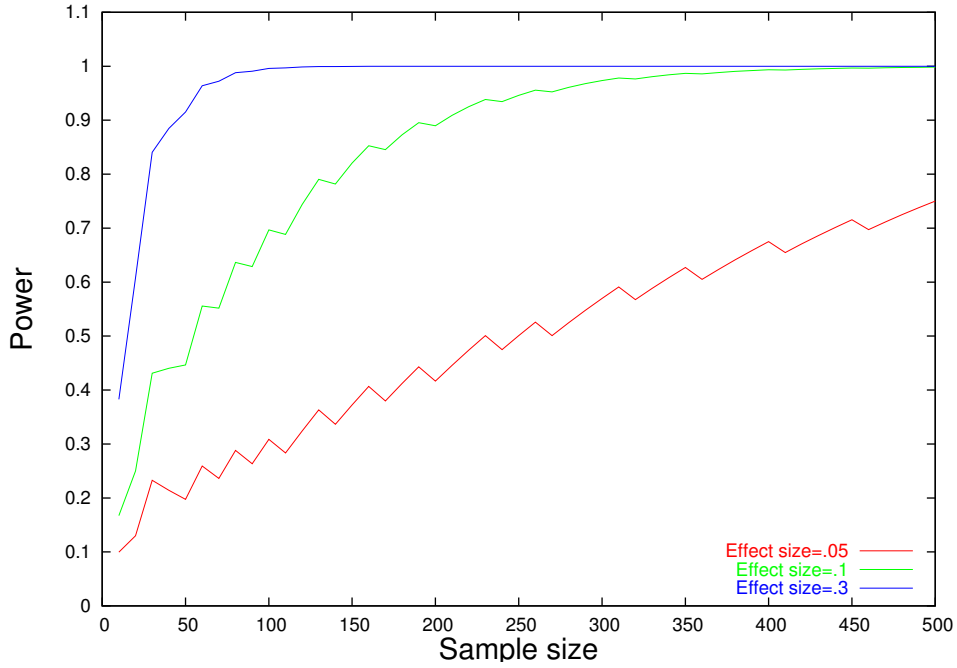
    In this case, for a sample size of 20, data generated under the alternative looks very different from data generated under the null (the probability of getting exactly 10/20 heads under the alternative is around 1 in 500,000).
- If the effect size is small, a large sample size is required to distinguish a data set generated by the null from a data set generated by the alternative. Consider the following two examples:

- Suppose the null hypothesis is $p = 1/2$ and the effect size is 0.01. If the sample size is one million and the null hypotehsis is true, with probability greater than 0.99 fewer than $501,500$ heads will be observed. If the alternative is true, with probability greater than 0.99 more than $508,500$ heads will be observed. Thus you are almost certain to identify the correct hypothesis based on such a large sample size.
- On the other hand, if the effect size is 0.4 (i.e. $p = 0.5$ vs. $p = 0.9$), under the null chances are greater than 97% that 14 or fewer heads will be observed in 20 flips. Under the alternative chances are greater than 98% that 15 or more heads will be observed in 20 flips. So only 20 observations are sufficient to have a very high chance of making the right decision in this case.

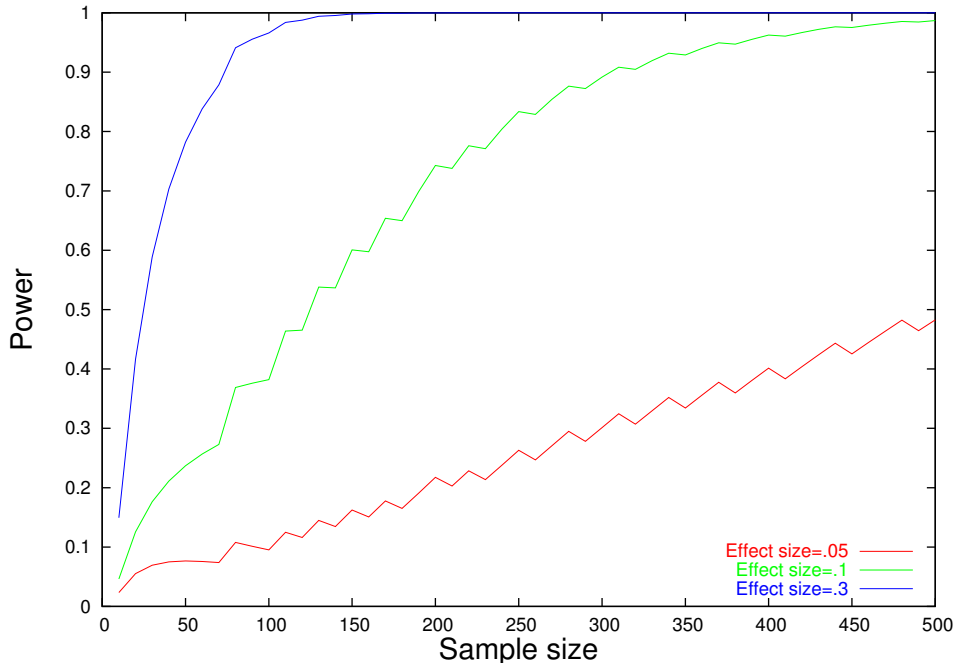
- To rationalize the trade-off between sample size and accuracy in hypothesis testing, it is common to calculate the power for various combinations of sample size and effect size. The power is the probability of observing a given level of evidence for the alternative when the alternative is true. Concretely, we may say that the power is the probability of observing a p-value smaller than .05 or .01 if the alternative is true.
- Usually the effect size is not known. However there are practical guidelines for establishing an effect size. Generally a very small effect is considered unimportant. For example, if patients treated under a new therapy survive less than one week longer on average compared to the old therapy, it may not be worth going to the trouble and expense of switching. Thus for purposes of planning an experiment, the effect size is usually taken to be the smallest difference that would lead to a change in practice.
- Once the effect size is fixed, the power can be calculated for a range of plausible sample sizes. Then power can be plotted against sample size.

  A plot of power against sample size always should have an increasing trend. However for technical reasons, the curve may sometimes drop slightly before resuming its climb.

- *Example:* For the one-sided coin flipping problem, suppose we would like to produce a p-value $< .05$ (when the alternative is true) for an effect size of .1, but we are willing to accept effect sizes as large as .3. The following figure shows power vs. sample size curves for effect sizes .1, .2, and .3.

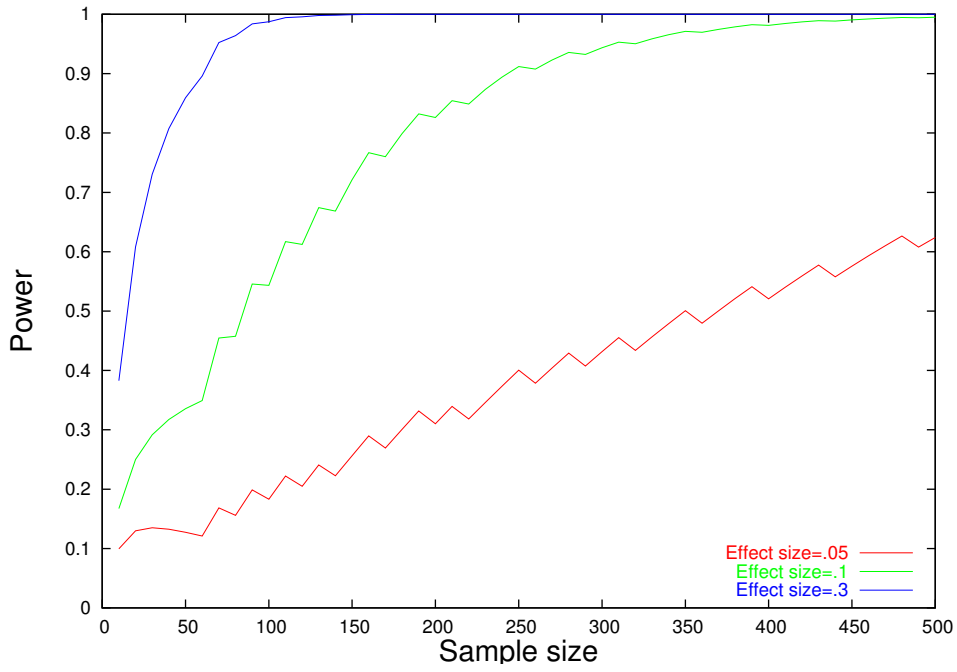


Power of obtaining p-value .05 vs. sample size for one-sided binomial test.

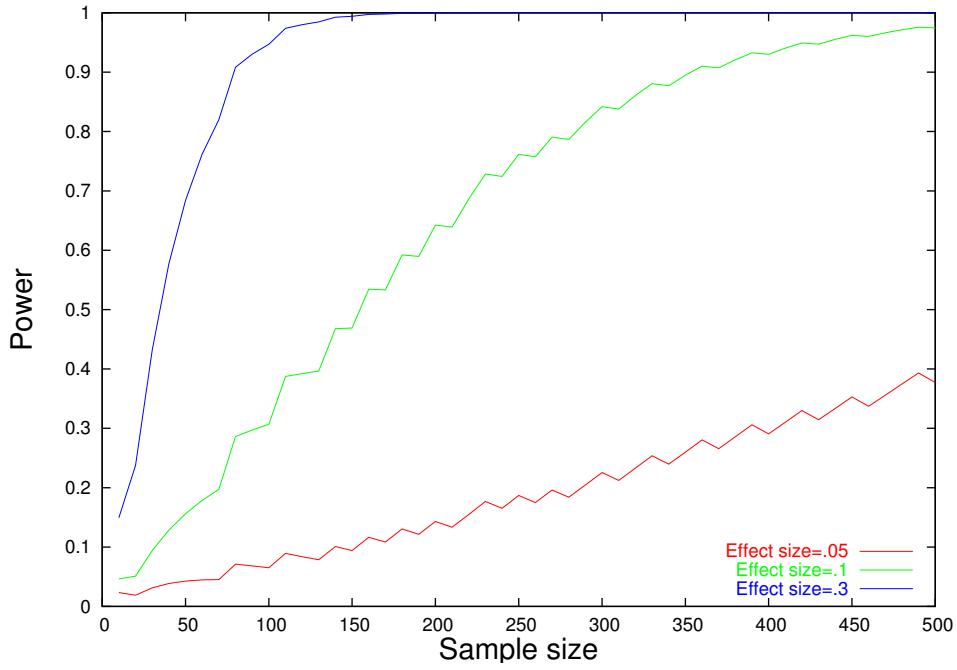


Power of obtaining p-value .01 vs. sample size for one-sided binomial test.

– *Example:* For the two-sided coin flipping problem, all p-values are twice correpsonding value in the one-sided problem. Thus it takes a larger sample size to achieve the same power.



Power of obtaining p-value .05 vs. sample size for two-sided binomial test.



Power of obtaining p-value .01 vs. sample size for two-sided binomial test.

– *Example:* Recall the BMI hypothesis testing problem from above. The test statistic was

$$T = (\bar{Y} - \bar{X})/\sqrt{\hat{\sigma}_X^2/m + \hat{\sigma}_Y^2/n}.$$

In order to calculate the p-value for a given value of T_{obs}, we need to know the distribution of T under the null hypothesis.

This can be done exactly, but for now we will accept as an approximation that $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ are exactly equal to the population values σ_X and σ_Y.

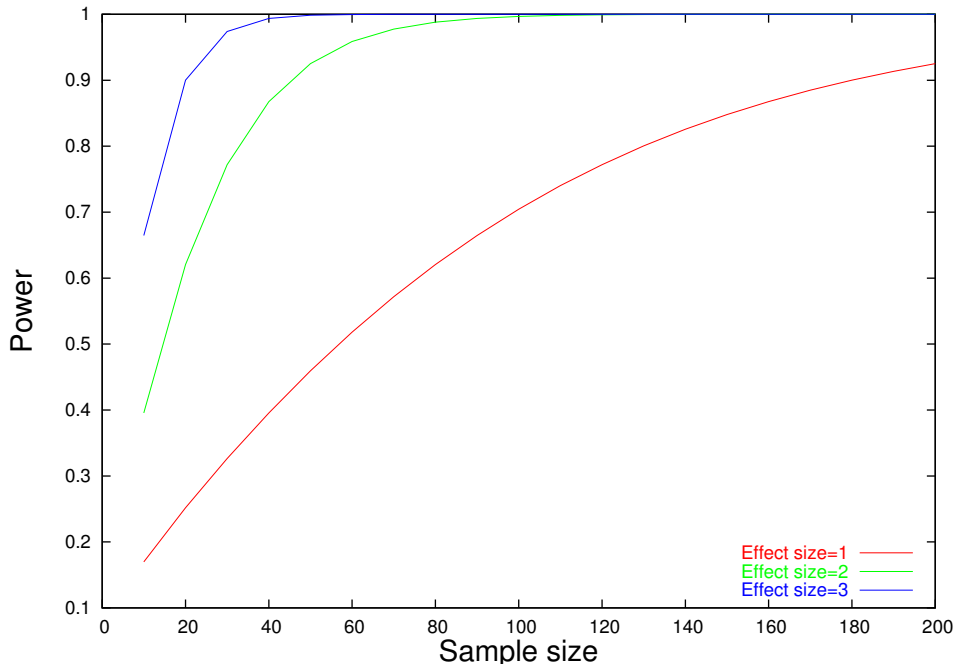
With this assumption, the expected value of T under the null hypothesis is 0, and its variance is 1. Thus we will use the standard normal distribution as an approximation for the distribution of T under the null hypothesis.
It follows that for the right-tailed test, T must exceed $Q(0.95) \approx 1.64$ to obtain a p-value less than 0.05, where Q is the standard normal quantile function.

Suppose that the Y (fast food eating) sample size is always 1/3 greater than the X (non fast food eating) sample size, so $n = 4m/3$. If the effect size is c (so $\mu_Y - \mu_X = c$), the test statistic can be written

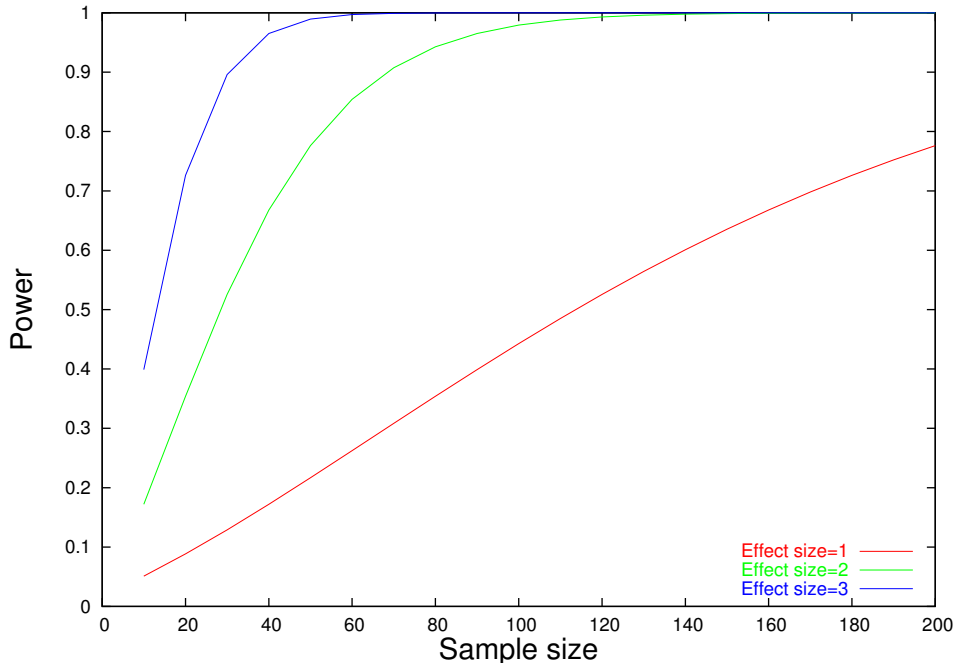
$$T = c/\hat{\sigma} + T^*, \quad T^* = (\bar{Y} - \bar{X} - c)/\hat{\sigma}$$

where $\hat{\sigma} = \sqrt{\hat{\sigma}_X^2/m + 3\hat{\sigma}_Y^2/(4m)}$ is the denominator of the test statistic.
Under the alternative hypothesis, T^* has mean 0 and standard deviation 1, so we will aprpoximate its distribution with a standard normal distribution.

Thus the power is $P(T > Q(.95)) = P(T^* > Q(.95) - c/\hat{\sigma})$, where probabilities are calculated under the alternative hypothesis. This is equal to $1 - F(Q(.95) - c/\hat{\sigma})$ (where F is the standard normal CDF). Note that this is a function of both c and m.



Power of obtaining p-value .05 vs. sample size for one sided Z-test.



Power of obtaining p-value .01 vs. sample size for one sided Z-test.

t-tests and Z-tests

- Previously we assumed that the estimated standard deviations $\hat{\sigma}_X$ and $\hat{\sigma}_Y$ were exactly equal to the population values σ_X and σ_Y. This allowed us to use the standard normal distribution to approximate p-values for the two sample Z test statistic:

$$(\bar{Y} - \bar{X})/\sqrt{\hat{\sigma}_X^2/m + \hat{\sigma}_Y^2/n}.$$

- The idea behind using the standard normal distribution here is:
  - The variance of $\bar{X}$ is σ_X^2/m and the variance of $\bar{Y}$ is σ_Y^2/n.
  - $\bar{X}$ and $\bar{Y}$ are independent, so the variance of $\bar{Y} - \bar{X}$ is the sum of the variance of $\bar{Y}$ and the variance of $\bar{X}$.

  Hence $\bar{Y} - \bar{X}$ has variance $\sigma_X^2/m + \sigma_Y^2/n$. Under the null hypothesis, $\bar{Y} - \bar{X}$ has mean zero. Thus

$$(\bar{Y} - \bar{X})/\sqrt{\hat{\sigma}_X^2/m + \hat{\sigma}_Y^2/n}.$$

  is approximately the "standardization" of $\bar{Y} - \bar{X}$.
- In truth,

$$\hat{\sigma}_X^2/m + \hat{\sigma}_Y^2/n$$

  and

$$\sigma_X^2/m + \sigma_Y^2/n$$

  differ somewhat, as the former is a random variable while the latter is a constant. Therefore, p-values calculated assuming that the Z-statistic is normal are slightly inaccurate.
- To get exact p-values, the following "two sample t-test statistic" can be used:

$$T = \sqrt{\frac{mn}{m+n}} \cdot \frac{\bar{Y} - \bar{X}}{S_p}$$

  where S_p^2 is the pooled variance estimate:

$$S_p^2 = \left((m-1)\hat{\sigma}_X^2 + (n-1)\hat{\sigma}_Y^2\right)/(m+n-2)$$

  The distribution of T under the null hypothesis is called t_{m+n-2}, or a "t distribution with $m+n-2$ degrees of freedom."

  p-values under a t distribution can be looked up in a table.

- *Example:* Suppose we observe the following:
  - $X_1, \ldots, X_{10}$, $\bar{X} = 1$, $\hat{\sigma}_X = 3$
  - $Y_1, \ldots, Y_8$, $\bar{Y} = 3$, $\hat{\sigma}_Y = 2$

  The Z test statistic is $(1-3)/\sqrt{9/10 + 1/2} \approx -1.6$, with a one-sided p-value of ≈ 0.05.

  The pooled variance is $S_p^2 = (9 \cdot 9 + 7 \cdot 4)/(10 + 8 - 2) \approx 6.8$ so $S_p \approx 2.6$. The two-sample t-test statistic is $\sqrt{80/18}(1-3)/2.6 \approx -1.62$, with $10+8-2 = 16$ df. The one-sided p-value is ≈ 0.06.
- The two sample Z or t-test is used to compare two samples from two populations, with the goal of inferring whether the two populations have the same mean.

  A related problem is to consider a sample from a single population, with the goal of inferring whether the population mean is equal to a fixed value, usually zero.
- Suppose we only have one sample $X_1, \ldots, X_n$ and we compute the sample mean $\bar{X}$ and sample standard deviation $\hat{\sigma}$. Then we can use

$$T = \sqrt{n}(\bar{X} - \theta)/\hat{\sigma}$$

  as a test statistic for the null hypothesis $\mu = \theta$ (where μ is the population mean of the X_i). Under the null hypothesis, T follows a t-distribution with $n-1$ degrees of freedom.

  Most often the null hypothesis is $\theta = 0$.
- For example, suppose we wish to test the null hypothesis $\mu = 0$ against an alternative $\mu > 0$. The test statistic is

$$T = \sqrt{n}\bar{X}/\hat{\sigma}.$$

  Under the null hypothesis T has a t_{n-1} distribution, which can be used to calculate p-values exactly. For example, if $\bar{X} = 6$, $n = 11$, and $\hat{\sigma} = 10$, then

$$T_{\text{obs}} = \sqrt{11} \cdot 3/5 \approx 2$$

  has a t_{10} distribution, which gives a p-value of around .04.
- If we use the same test statistic as above, but assume that $\hat{\sigma} = \sigma$, then we can use the normal approximation to get an approximate p-value.

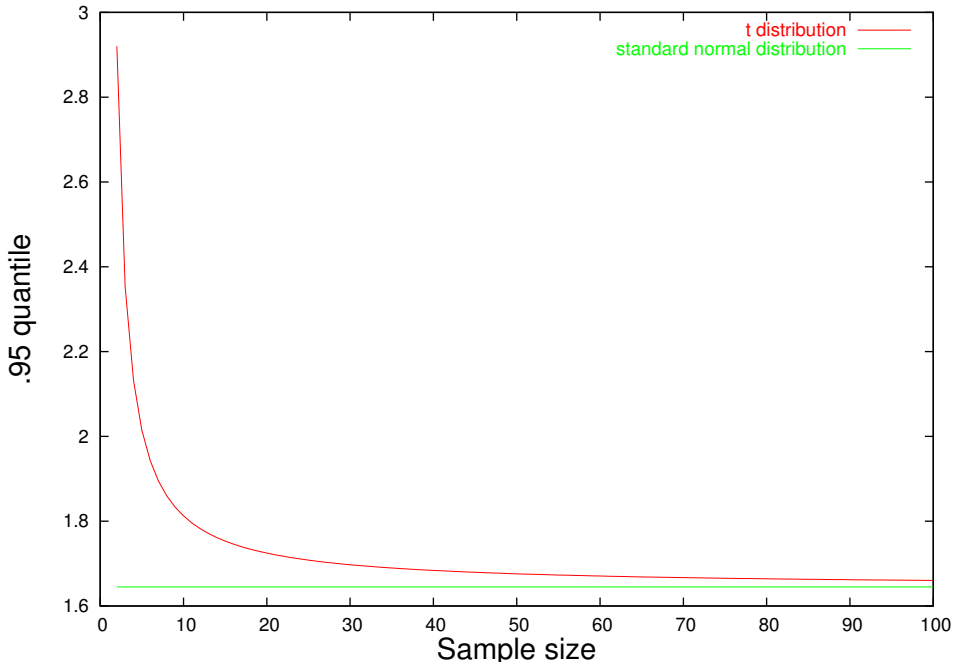
  For the example above, the Z statistic p-value is .02 which gives an overly strong assessment of the evidence for the alternative compared to the exact p-value computed under the t distribution.

  If we were to use the two sided alternative $\mu \neq 0$, then the p-value would be .07 under the t_{10} distribution and .05 under the standard normal distribution.

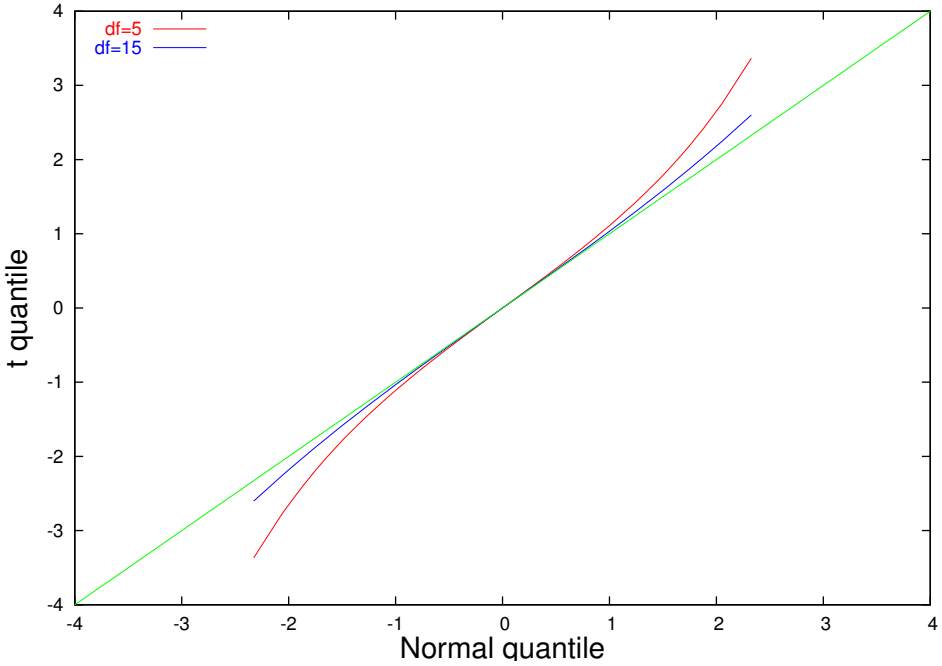
- For small degrees of freedom, the t distribution is substantially more variable than the standard normal distribution.

  Therefore under a t-distribution the p-values will be somewhat larger (suggesting less evidence for the alternative).

  If the sample size is larger than 50 or so, the two distributions are so close that they can be used interchangeably.



.95 quantile for the t-distribution as a function of sample size, and the .95 quantile for the standard normal distribution.



QQ plot comparing the quantiles of a standard normal distribution (x axis) to the quantiles of the t-distribution with two different degrees of freedom.

- A special case of the one-sample test is the paired two-sample test. Suppose we make observations X_1, Y_1 on subject 1, X_2, Y_2 on subject 2, etc. For example, the observations might be "before" and "after" measurements of the same quantity (e.g. tumor size before and after treatment with a drug).

  Let $D_i = Y_i - X_i$ be the change for subject i. Now suppose we wish to test whether the before and after measurements for each subject have the same mean. To accomplish this we can do a one-sample Z-test or t-test on the D_i.

  If the data are paired, it is much better to do a paired test, rather than to ignore the pairing and do an unpaired two-sample test. We will see why this is so later.
- *Example:* Suppose we observe the following paired data:

X	Y	D
5	4	1
2	1	1
9	7	2

X	Y	D
7	3	2
6	5	1
3	1	2

  $\bar{D} = 1.5$ and $\hat{\sigma}_D = \sqrt{0.3}$, so the paired test statistic is $\sqrt{6} \cdot 1.5/\sqrt{0.3} \approx 16$, which is highly significant.

  $\bar{X} = 16/3$, $\hat{\sigma}_X \approx 2.6$, $\bar{Y} = 21/6$, $\hat{\sigma}_Y \approx 2.3$, so the unpaired two-sample Z test statistic is $(16/3 - 21/6)/\sqrt{2.6^2/6 + 2.3^2/6} \approx 0.9$ which is not significant.

- The one and two sample t-statistics only have a t-distribution when the underlying data have a normal distribution.

  Moreover, for the two sample test the population standard deviations σ_X and σ_Y must be equal.

  If the sample size is large, then p-values computed from the standard normal or t-distributions will not be too far from the true values even if the underlying data are not normal, or if σ_X and σ_Y differ.

Summary of One and Two Sample Tests

	Test statistic	Reference distribution
One sample Z	$\sqrt{m} \cdot \bar{X}/\hat{\sigma}_X$	N(0,1)
Paired Z	$\sqrt{m} \cdot \bar{D}/\hat{\sigma}_D$	N(0,1)
Two sample Z	$(\bar{Y} - \bar{X})/\hat{\sigma}_{XY}$	N(0,1)
One sample t	$\sqrt{m} \cdot \bar{X}/\hat{\sigma}_X$	t_{m-1}
Paired t	$\sqrt{m} \cdot \bar{D}/\hat{\sigma}_D$	t_{m-1}
Two sample t	$\sqrt{\frac{mn}{m+n}} \cdot \frac{\bar{Y}-\bar{X}}{S_p}$	t_{m+n-2}

$\hat{\sigma}^2_{XY} = \hat{\sigma}^2_X/m + \hat{\sigma}^2_Y/n$

$S_p^2 = (\sum(Y_i - \bar{Y})^2 + \sum(X_i - \bar{X})^2)/(m + n - 2)$

Confidence intervals and prediction intervals

- Suppose that our goal is to estimate an unknown constant. For example, we may be interested in estimating the acceleration due to gravity g (which is $9.8m/s^2$).

  We assume that our experimental measurements are unbiased, meaning that the mean of each X_i is g. In this case, it makes sense to estimate g using $\bar{X}$.

  The value of $\bar{X}$ is a point estimate of g. But we would like to quantify the uncertainty in the estimate.
- In general, suppose we are using $\bar{X}$ as a point estimate for an unknown constant θ. The estimation error is $\bar{X} - \theta$.

  For a given value of $c > 0$, we can calculate the probability that the estimation error is smaller than c: $P(|\bar{X} - \theta| \leq c)$.

  Standardizing $\bar{X}$ yields $\sqrt{n}(\bar{X}-\theta)/\hat{\sigma}$, which has a t-distribution with $n-1$ degrees of freedom (assuming that the measurements are normal). Thus

$$\begin{aligned} P(|\bar{X} - \theta| \leq c) &= P(\sqrt{n}|\bar{X} - \theta|/\hat{\sigma} \leq \sqrt{n}c/\hat{\sigma}) \\ &= P(T \leq \sqrt{n}c/\hat{\sigma}) \end{aligned}$$

  which can be determined from a table of the t_{n-1} distribution.

  This quantity is called the coverage probability.
- We would like to control the coverage probability by holding it at a fixed value, usually 0.9, 0.95, or 0.99.

  This means we will set

$$\sqrt{n}c/\hat{\sigma} = Q,$$

  where Q is the $1 - (1 - \alpha)/2$ quantile of the t_{n-1} distribution for $\alpha = 0.9, 0.95$, etc. Solving this for c yields

$$c = Q\hat{\sigma}/\sqrt{n}.$$

- Thus the $100\times\alpha\%$ confidence interval (CI) is

$$\bar{X} \pm Q\hat{\sigma}/\sqrt{n},$$

  which may also be written

$$(\bar{X} - Q\hat{\sigma}/\sqrt{n}, \bar{X} + Q\hat{\sigma}/\sqrt{n}).$$

  The width of the CI is $2Q\hat{\sigma}/\sqrt{n}$. Note how its scales with α, $\hat{\sigma}$, and n.

  If n is not too small, normal quantiles can be used in place of t_{n-1} quantiles.

- *Example:* Suppose we observe $X_1, \ldots, X_{10}$ with $\bar{X} = 9.6$ and $\hat{\sigma} = 0.7$. The CI is

$$9.6 \pm 2.26 \cdot 0.7/\sqrt{10},$$

or 9.6 ± 0.5.

- A terminological nuance for a 95% CI:

OK: "There is a 95% chance that θ falls within $\bar{X} \pm Q\hat{\sigma}/\sqrt{n}$.

Better: "There is a 95% chance that the interval $\bar{X} \pm Q\hat{\sigma}/\sqrt{n}$ covers θ.

- Whether a confidence interval is a truthful description of the actual error distribution may depend strongly on the assumption that the data (i.e. the X_i) follow a normal distribution. If the data are strongly non-normal (e.g. skewed or with thick tails), the CI is typically inaccurate (i.e. you tell somebody that a CI has a 95% chance of containing the true value, but the actual probability is lower).
- Confidence intervals are often reported casually as "margins of error". For example you may read in the newspaper that the proportion of people supporting a certain government policy is $.7 \pm .03$. This statement doesn't mean anything unless the probability is given as well. Generally, intervals reported this way in newspapers, etc., are 95% CI's, but the 95% figure is almost never stated.
- Suppose we wish to quantify the uncertainty in a prediction that we make of a future observation. For example, today we observe a SRS $X_1, \ldots, X_n$ and tomorrow we will observe a single additional observation Z from the same distribution.

Concretely, we may be carrying out a chemical synthesis in which fixed amounts of two reactants are combined to yield a product. Our goal is to predict the value of Z before observing it, and to quantify the uncertainty in our prediction.

Since the X_i and Z have the same mean, our prediction of Z will be $\bar{X}$. In order to quantify the prediction error we will find c so that $P(|Z - \bar{X}| \leq c) = \alpha$. This is called the $100 \cdot \alpha\%$ prediction interval.

To find c, note that $Z - \bar{X}$ has mean 0 and standard deviation $\sqrt{(n+1)\sigma^2/n}$. Thus

$$\begin{aligned} P(|Z - \bar{X}| \leq c) &= P(\sqrt{n}|Z - \bar{X}|/\hat{\sigma}\sqrt{n+1} \leq \sqrt{n}c/\hat{\sigma}\sqrt{n+1}) \\ &= P(T \leq \sqrt{n}c/\hat{\sigma}\sqrt{n+1}) \end{aligned}$$

Let Q be the $1 - (1-\alpha)/2$ quantile of a t_{n-1} distribution. Solving for c yields

$$c = Q\sqrt{\frac{n+1}{n}}\hat{\sigma}.$$

Note how this scales with Q, n, and σ.

If n is not too small, normal quantiles can be substituted for the t_{n-1} quantiles.

- *Example:* Suppose that $n = 8$ replicates of an experiment carried out today yielded $\bar{X} = 17$ and $\hat{\sigma} = 1.5$. The 95% PI is

$$\bar{X} \pm 2.36 \cdot \sqrt{9/8} \cdot 1.5,$$

  or 17 ± 3.8.
- The width of the CI goes to zero as the sample size gets large, but the width of the PI never is smaller than $2Q\sigma$.

  The CI measures uncertainty in a point estimate of an unknown constant. This uncertainty arises from estimation of μ (using $\bar{X}$) and σ (using $\hat{\sigma}$).

  The PI measures uncertainty in an unobserved random quantity. This includes the uncertainty in $\bar{X}$ and $\hat{\sigma}$, in addition to uncertainty in Z. This is why the PI is wider.
- **Summary of CI's and PI's:** To obtain a CI or PI with coverage probability α for a SRS $X_1 \ldots, X_n$ with sample mean $\bar{X}$ and sample standard deviation $\hat{\sigma}$, let $\alpha^* = 1 - (1-q)/2$, let Q_N be the standard normal quantile function, and let Q_T be the t_{n-1} distribution quantile function.

	CI	PI
Approximate	$\bar{X} \pm Q_N(\alpha^*)\hat{\sigma}/\sqrt{n}$	$\bar{X} \pm \sqrt{(n+1)/n} Q_N(\alpha^*)\hat{\sigma}$
Exact	$\bar{X} \pm Q_T(\alpha^*)\hat{\sigma}/\sqrt{n}$	$\bar{X} \pm \sqrt{(n+1)/n} Q_T(\alpha^*)\hat{\sigma}$

Transformations

- The accuracy of confidence intervals, prediction intervals, and p-values may depend strongly upon whether the data follow a normal distribution.

  Normality is critical if the sample size is small, but much less so for large sample sizes.

  It is a good idea to check the normality of the data before giving too much credence to the results of any statistical analysis that depends on normality.

  The main diagnostic for assessing the normality of a SRS is the normal probability plot.
- If the data are not approximately normal, it may be possible to transform the data so that they become so.

  The most common transformations are

$$\begin{array}{lcll} X_i & \rightarrow & \log(X_i - c) & \text{logarithmic transform} \\ X_i & \rightarrow & (X_i - c)^q & \text{power transform} \\ X_i & \rightarrow & \log(X_i - c + \sqrt{(X_i - c)^2 + d}) & \text{generalized log transform} \\ X_i & \rightarrow & \log(X_i/(1 - X_i)) & \text{logistic transform} \end{array}$$

  We select $c > \min X_i$ to ensure that the transforms are defined. The logistic transform is only applied if $0 \leq X_i \leq 1$.

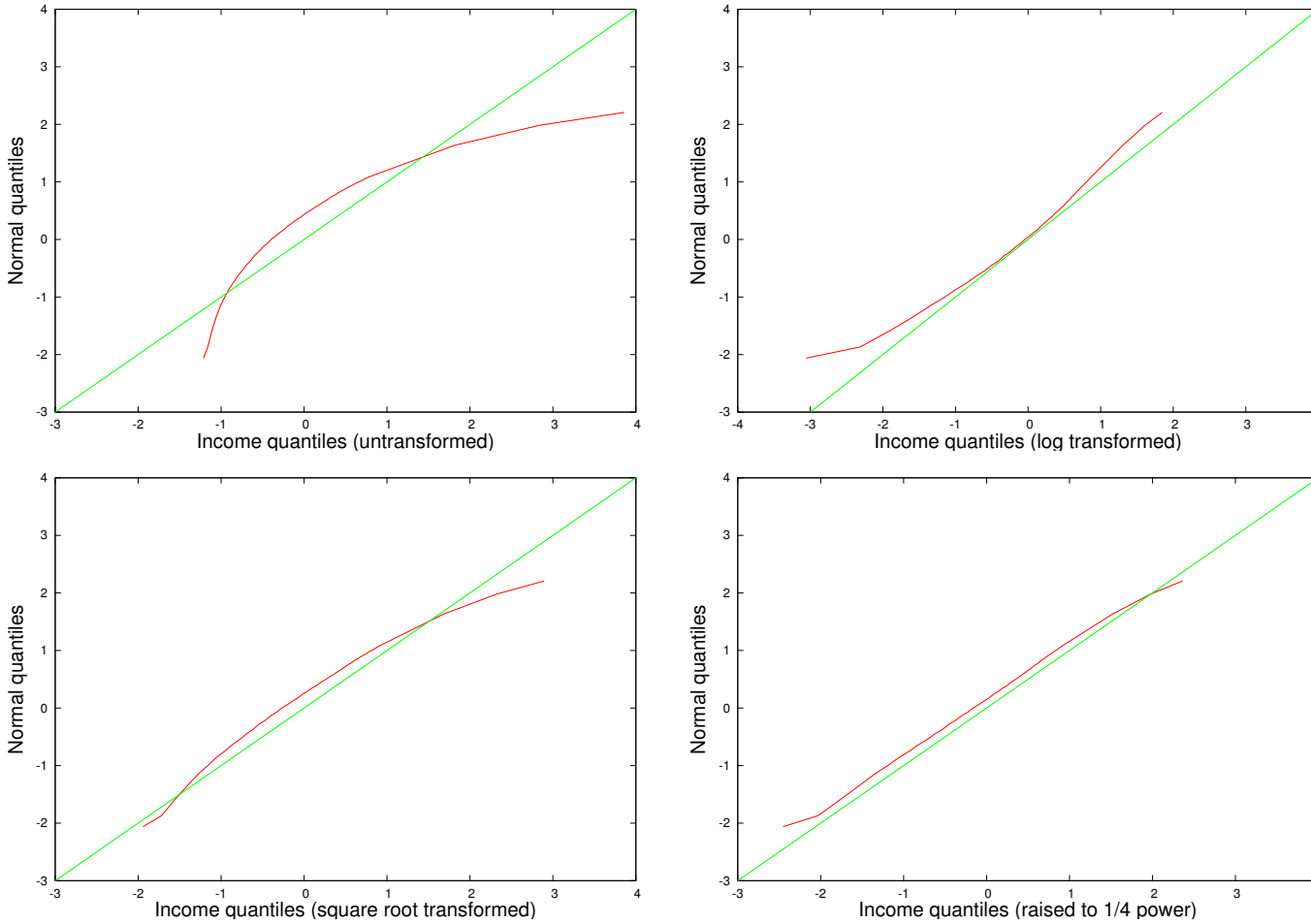
  The values c, q, and d are chosen to improve the normality of the data.
- As $q \rightarrow 0$ the power transform becomes more like a log-transform:

$$\frac{d}{dx} \log x = 1/x \qquad \frac{d}{dx} x^q \propto x^{q-1}$$

$$\frac{d^2}{dx^2} \log x = -1/x^2 \qquad \frac{d^2}{dx^2} x^q \propto x^{q-2}$$

- Log transforms and power transforms (with $q < 1$) generally are used to reduce right skew. The log transform carries out the strongest correction, while power transforms are milder.
- The upper left plot in the following figure shows a normal probability plot for the distribution of US family income in 2001. The rest of the figure contains normal probability plots for various transformations of the income data.

  The transform $X \rightarrow X^{1/4}$ is the most effective at producing normality.



- The log transform is especially common for right skewed data because the transformed values are easily interpretable. With log-transformed data, differences become fold changes. For example, if $X_i^* = \log(X_i)$ and $Y_i^* = \log(Y_i)$, then

$$\begin{aligned}
\bar{X}^* - \bar{Y}^* &= \sum_i \log(X_i)/m - \sum_j \log(Y_j)/n \\
&= \log(\prod_i X_i)/m - \log(\prod_j Y_j)/n \\
&= \log\left((\prod_i X_i)^{1/m}\right) - \log\left((\prod_j Y_j)^{1/n}\right) \\
&= \log\left((\prod X_i)^{1/m}/(\prod Y_j)^{1/n}\right),
\end{aligned}$$

where $(\prod X_i)^{1/m}$ and $(\prod Y_j)^{1/n}$ are the geometric means of the X_i and the Y_i respectively.

- If base 10 log are used, and $\bar{X}^* - \bar{Y}^* = c$, then we can say that the X values are 10^c times greater than the Y values on average (where "average" is the "geometric average", to be precise).

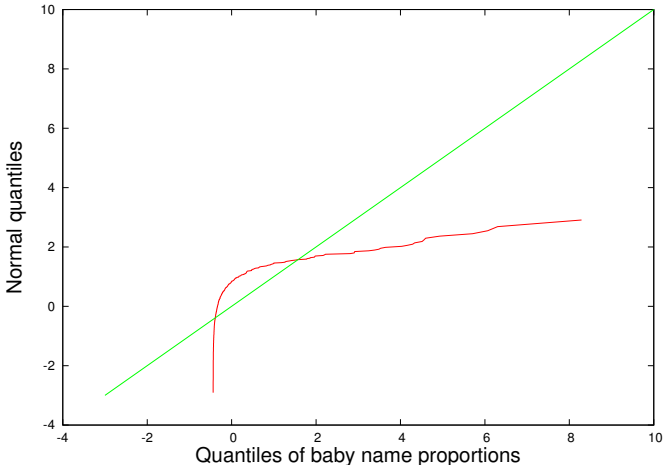
  Similarly, if base 2 logs are used, we can say that the X values are 2^c times greater than the Y values on average, or there are "c doublings" between the X and Y values (on average).
- The normal probability transform forces the normal probability plot to follow the diagonal exactly. If the sample size is n, and Q is the normal quantile function, we have:

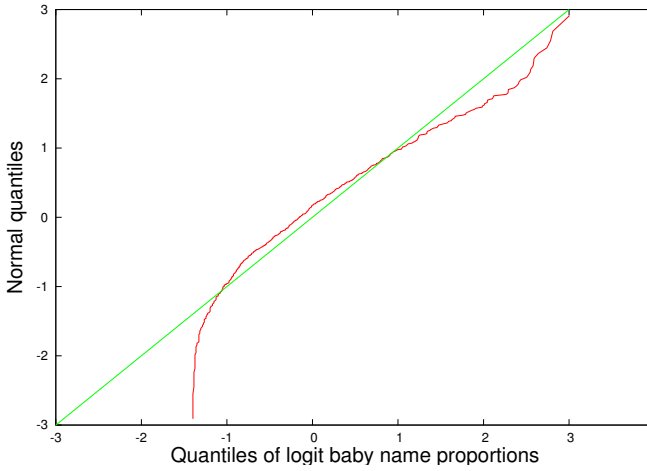
$$X_{(k)} \to Q(k/(n+1)).$$

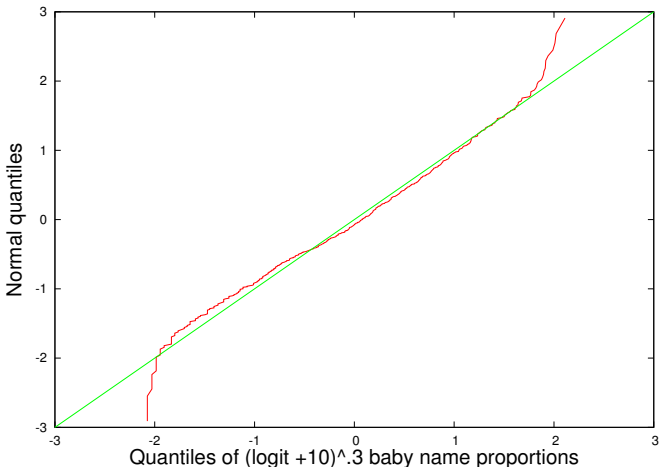
  The main drawback to this transform is that it is difficult to interpret or explain what has been done.
- Data occuring as proportions are often not normal. To improve the normality of this type of data, apply the logistic transform. This transform maps $(0, 1)$ to $(-\infty, \infty)$.
- The following figures show normal probability plots for the proportion of male babies in the 1990's given each of the 500 most popular names. QQ plots are shown for unstransformed data, for logit transformed data, and for data transformed via $X \to (\text{logit}(X) + 10)^{3/10}$.

  The unstransformed data are seen to be strongly non-normal (right skewed). The logit scale data are much better, but still show substantial deviation from normality (the left tail that drops below the diagonal comprises around 17% of the data).

  The transform $(\text{logit}(X) + 10)^{3/10}$ brings the distribution very close to normality (there are deviations in the extreme tails, but these account for less than 2% of the data).


Normal quantiles
Quantiles of baby name proportions


Normal quantiles
Quantiles of logit baby name proportions


Normal quantiles
Quantiles of (logit +10)^.3 baby name proportions

Multiple testing

- Under the null hypothesis, the probability of observing a p-value smaller than p is equal to p. For example, the probability of observing a p-value smaller than .05 is .05.

  It is often the case that many hypotheses are being considered simultaneously. These are called simultaneous hypotheses.

  For each hypothesis individually, the chance of making a false positive decision is .05, but the chance of making a false positive decision for at least one of several hypotheses is much higher.

- *Example:* Suppose that the IRS has devised a test to determine if somebody has cheated on his or her taxes. A test statistic T is constructed based on the data in a tax return, and a critical point T_{crit} is determined such that $T \geq T_{\text{crit}}$ implies a p-value of less than .01.

  Suppose that 100 tax returns are selected, and that in truth nobody is cheating (so all 100 null hypotheses are true). Let $X_i = 1$ if the test for return i yields a p-value smaller than .01, let $X_i = 0$ otherwise, and let $S = X_1 + \cdots + X_n$ (the number of accusations).

  The distribution of S is binomial with $n = 100$ and success probability $p = .01$. The mean of S is 1, and $P(S = 0) = (99/100)^{100} \approx .37$. Thus chances are around 2/3 that somebody will be falsely accused of cheating. If $n = 500$ then the chances are greater than 99%.

- The key difficulty in screening problems (like the tax problem) is that we have no idea which person to focus on until after the data are collected.

  If we have reason to suspect a particular person ahead of time, and if the p-value for that person's return is smaller than .01, then the chances of making a false accusation are .01. The problem only arises when we search through many candidate hypotheses to find the one with the strongest evidence for the alternative.

- One remedy is to require stronger evidence (i.e. a smaller p-value) for each individual test. If the p-values are required to be small enough, the overall probability of making a false positive can be made smaller than any given level.

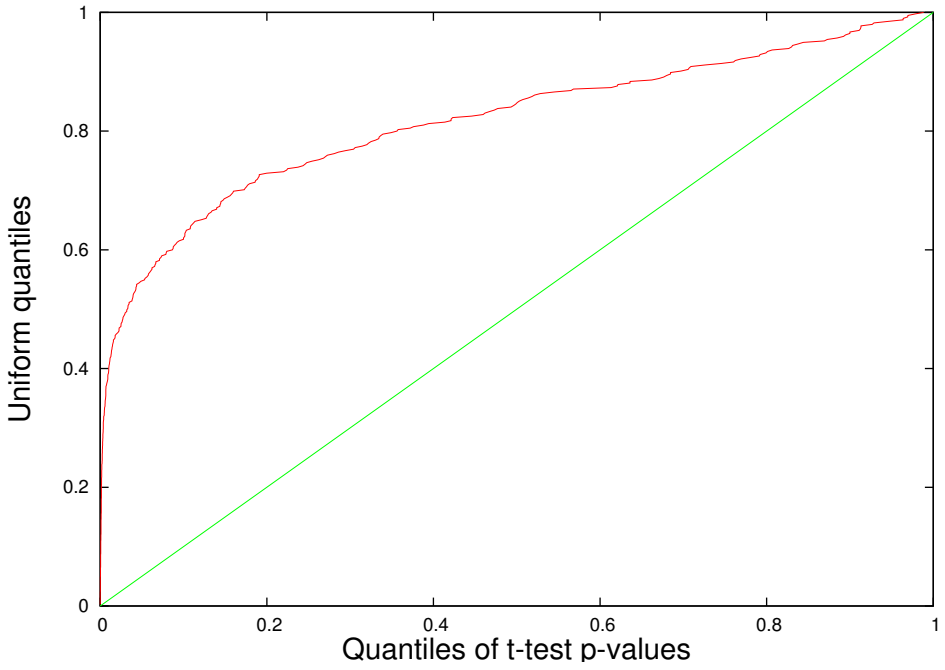
  In the above example, the p-value for each test must satisfy $1 - (1 - p)^n = .05$, giving $p \approx .0005$ when $n = 100$. Unfortunately, if we require this level of evidence we will have very little power.

- A different remedy is to change the goal – instead of attempting to acquire strong evidence that a specific person is cheating, we aim to acquire strong evidence that cheating is taking place somewhere in the population. This would be helpful in determining whether tax policy should be changed, but it would not allow us to prosecute anybody for current transgressions.

  Recall that out of a large number of hypothesis tests for which the null hypothesis is true, on average fraction p of the tests will yield p-values smaller than p. Thus we can form a QQ plot comparing the uniform quantiles $Q(k/n) = k/n$ (which are the correct quantiles if the null hypothesis is always true) to the observed quantiles for the p-values $p_1, \ldots, p_n$ in all the hypothesis tests.

  If the QQ shows that the p-values are strongly left-skewed compared to a uniform distribution, there is substantial evidence in the data that some people are cheating, even though we don't have sufficient evidence to prosecute any specific people.

- As an example, for each decade over the last century, the US Social Security Administration calculated the proportion of baby boys given each of the 1000 most popular names in a given decade (based on a sample of 5% of all social security registrations). We can extract from these names the 395 names that occur at least once in each decade of the twentieth century, and convert their proportions using the transform $X \to (\text{logit}(X) + 10)^{3/10}$ described above. Using two sample t tests, for each name we can test whether the mean proportion in the first five decades (1900-1949) is equal to the mean proportion in the second five decades (1950-2000). Since we are carrying out a hypothesis test for each name, we are carrying out 395 hypothesis tests. The QQ plot of the t test p-values follows.



Since the quantiles corresponding to the t test p-values are much smaller than the corresponding uniform quantiles, we conclude that there is substantial evidence that at least some names became more popular or less popular in the second half of the century.

This does not quantify the evidence for specific names that may have changed, but we can look at the smallest individual p-values to determine some of the best candidates ("Conrad" down by a factor of 3, "Dexter" up by a factor of 10, "Christopher" up by a factor of 100).