University of Washington

STATISTICS

Applied Statistics and Experimental Design

Normal and Related Distributions (t, χ^2 and F)

Tests, Power, Sample Size & Confidence Intervals

Fritz Scholz

Fall Quarter 2008

Hypothesis Testing

We have addressed the question: Does the type of flux affect SIR?

Formally we have tested the

null hypothesis H_0: The type of flux does not affect SIR

against the

alternative hypothesis H_1: The type of flux does affect SIR.

While H_0 seems fairly specific, H_1 is open ended. H_1 can be anything but H_0.

There may be many ways for SIR to be affected by flux differences,

e.g., change in mean, median, or scatter. Different effects for different boards?

Such differences may show up in the data vector **Z** through an appropriate test statistic $s(\mathbf{Z})$. Here $\mathbf{Z} = (X_1, \ldots, X_9, Y_1, \ldots, Y_9)$.

Test Criteria or Test Statistics

In the flux analysis we chose to use the absolute difference of sample means, $s(\mathbf{Z}) = |\bar{Y} - \bar{X}|$, as our test criterion or test statistic for testing the null hypothesis.

A test statistic is a value calculated from data and other known entities, e.g., assumed (e.g., hypothesized) parameter values.

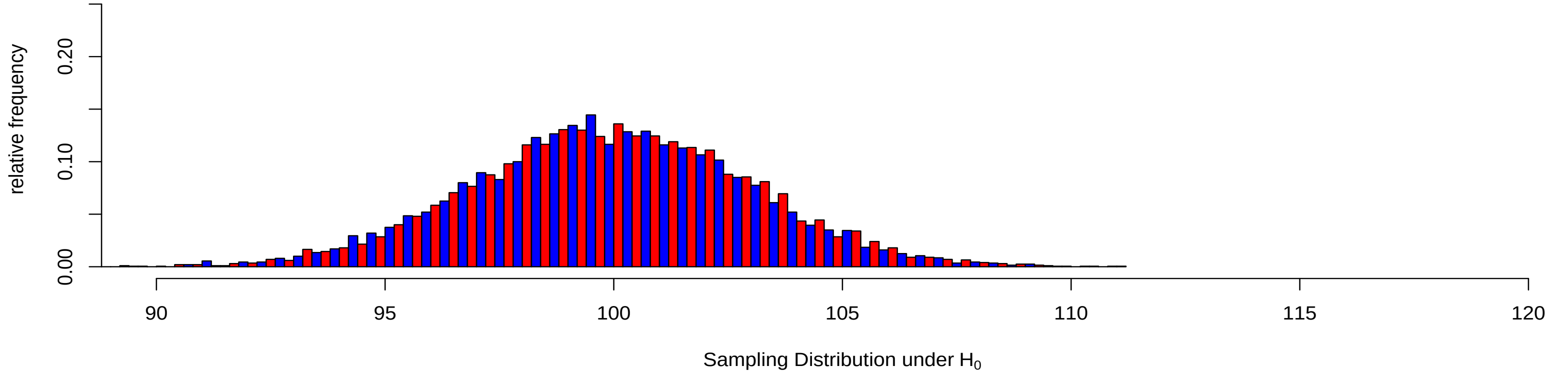
We could have worked with the absolute difference in sample medians or with the ratio of sample standard deviations and compared that ratio with 1, etc.

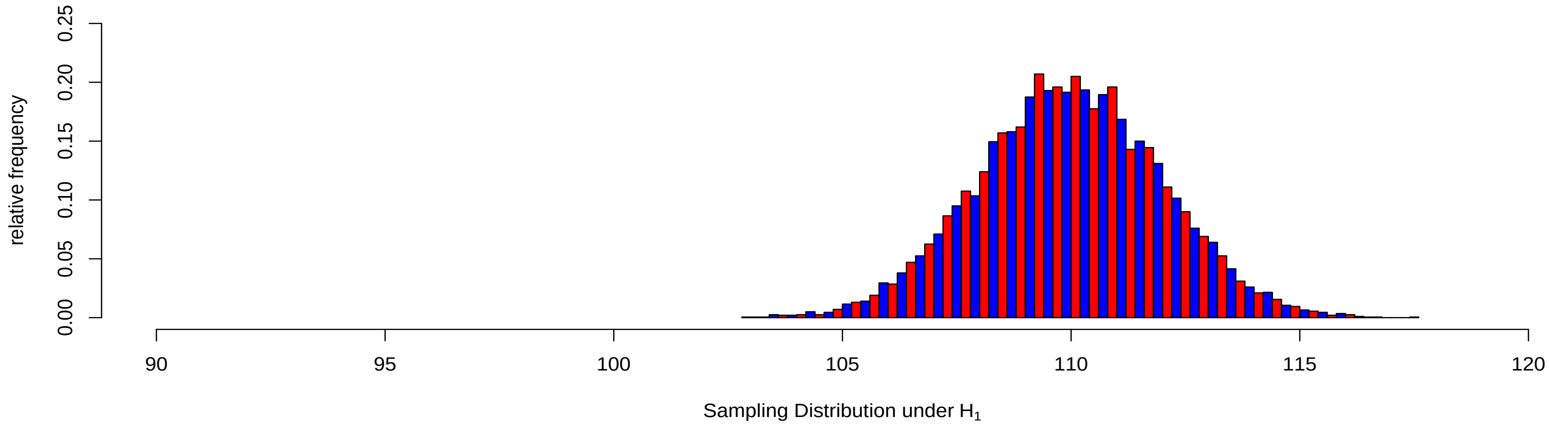
Different test statistics are sensitive to different deviations from the null hypothesis.

A test statistic, when viewed as a function of random input data, is itself a random variable, and has a distribution, its sampling distribution.

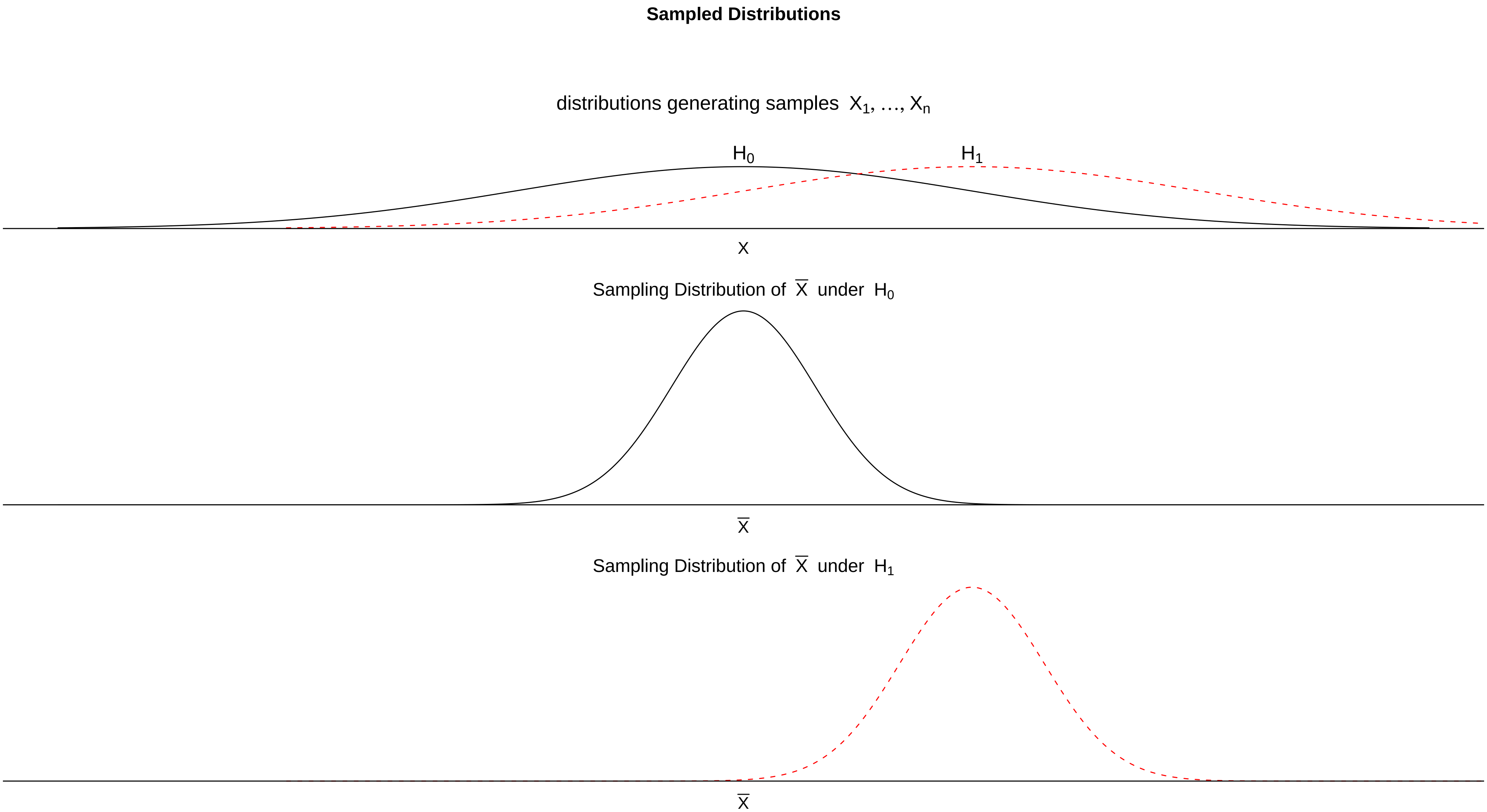
Sampling Distributions

For a test statistic $s(\mathbf{Z})$ to be effective in deciding between H_0 and H_1, the sampling distributions of $s(\mathbf{Z})$ under H_0 and H_1 should be separable to some degree.





Sampled and Sampling Distributions



When to Reject H_0

The previous illustration shows a specific sampling distribution for $s(\mathbf{Z})$ under H_1.

Typically H_1 consists of many different possible distributional models leading to many possible sampling distributions under H_1.

Under H_0 we often have just a single sampling distribution, the null distribution.

If under H_1 the test statistics $s(\mathbf{Z})$ tends to have mostly higher values than under H_0, we would want to reject H_0 when $s(\mathbf{Z})$ is large, as on the previous slide.

How large is too large? Need a critical value C_{crit} and reject H_0 when $s(\mathbf{Z}) \geq C_{\text{crit}}$.

Choose C_{crit} such that $P(s(\mathbf{Z}) \geq C_{\text{crit}} | H_0) = \alpha$, a pre-chosen significance level. Typically $\alpha = .05$ or $.01$. It is the probability of the type I error.

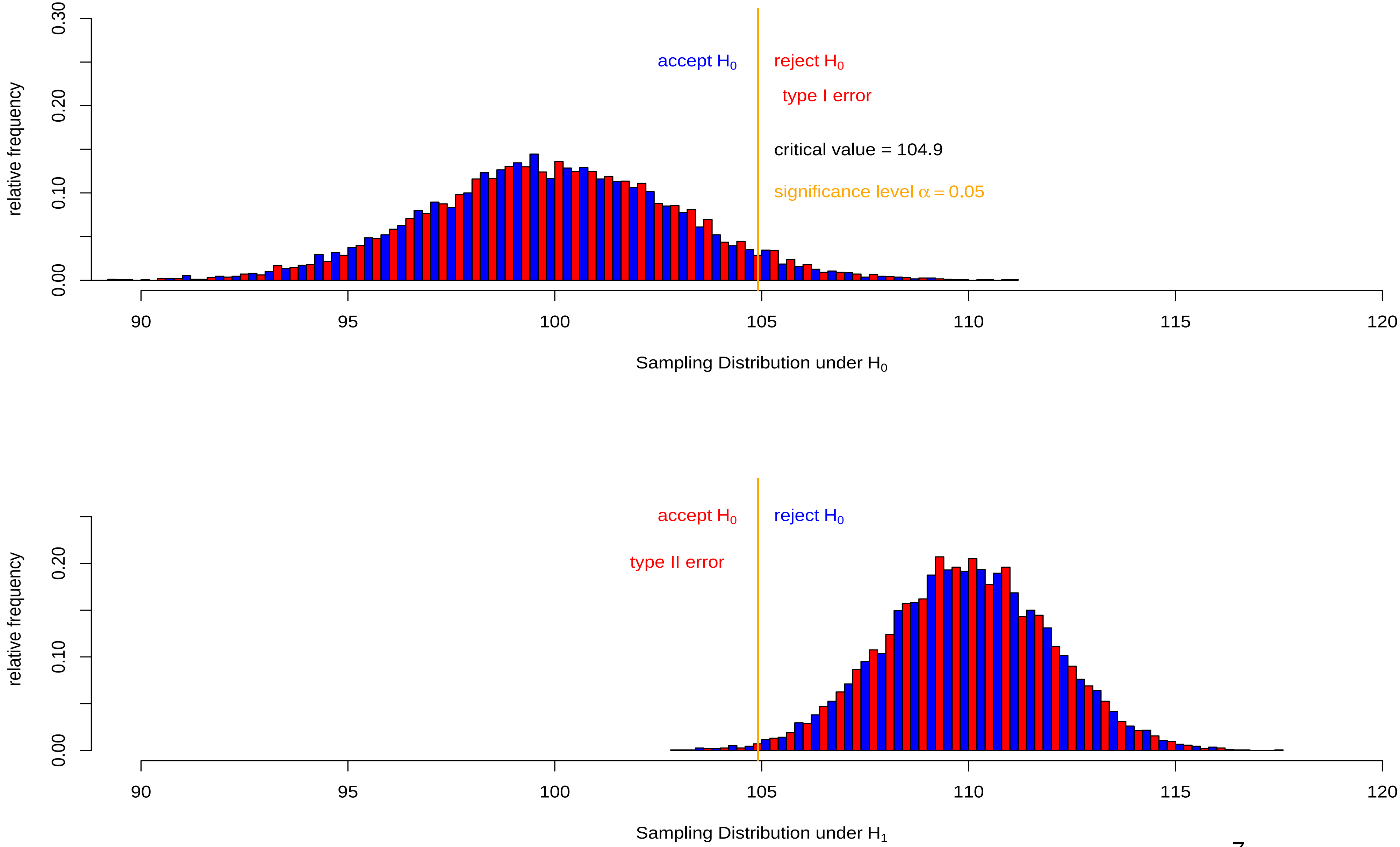
The previous illustration also shows that there may be values $s(\mathbf{Z})$ in the overlap of both distributions. Decisions are not clear cut $\Longrightarrow$ type I or type II error

Decision Table

	Truth	
Decision	H_0 is true	H_0 is false
accept H_0	correct decision	type II error
reject H_0	type I error	correct decision

Testing hypotheses (like estimation) is a branch of a more general framework , namely decision theory. Decisions are optimized with respect to penalties for wrong decisions, i.e., $P(\text{Type I Error})$ and $P(\text{Type II Error})$, or the mean squared error of an estimate $\hat{\theta}$ of θ, namely $E((\hat{\theta}-\theta)^2)$.

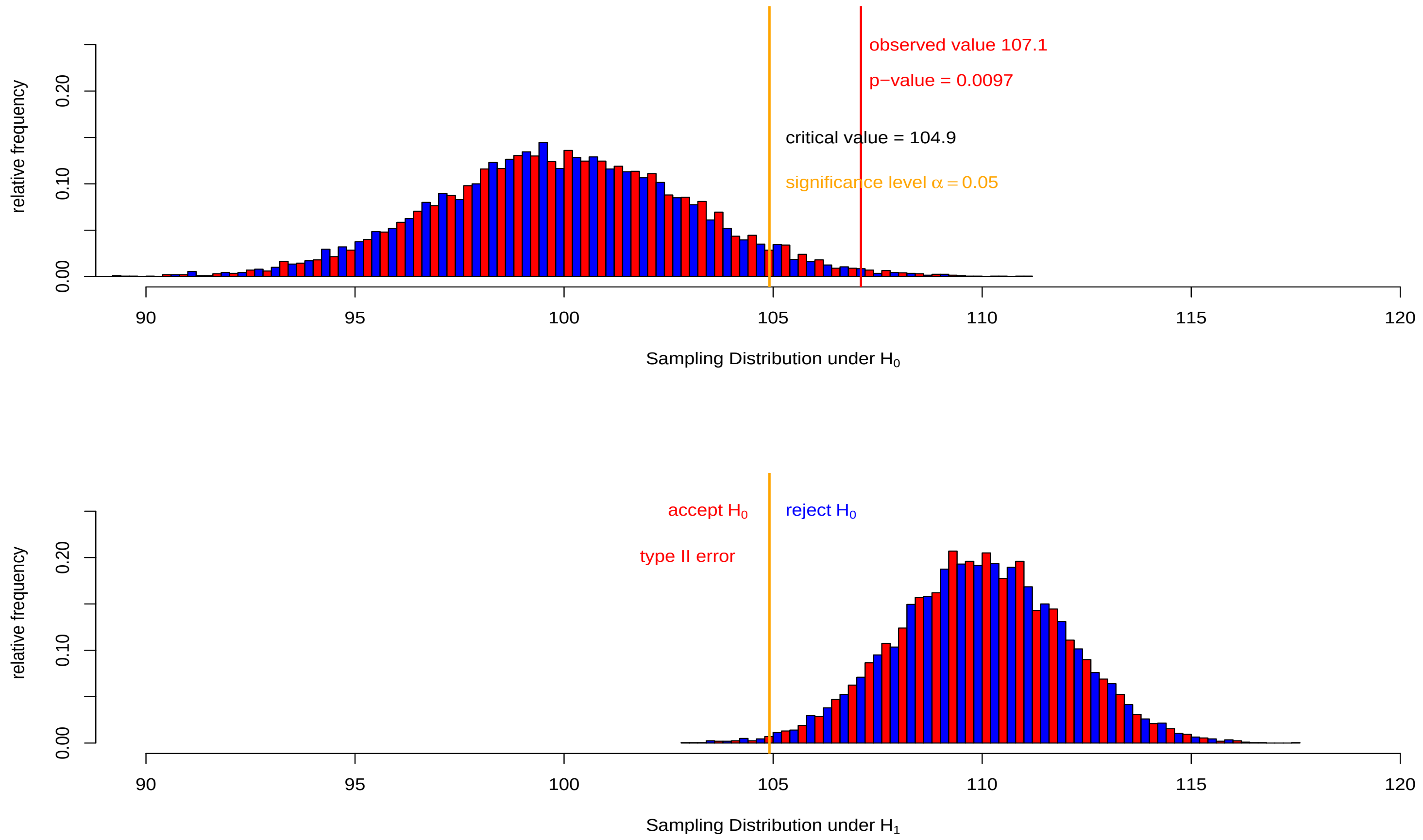
The Null Distribution and Critical Values



Critical Values and p-Values

The p-value$(s(\mathbf{z}))$ for the observed test statistic $s(\mathbf{z})$ is $P(s(\mathbf{Z}) \geq s(\mathbf{z})|H_0)$.

Note that p-value$(s(\mathbf{z})) \leq \alpha$ is equivalent to rejecting H_0 at level α.



p-Values and Significance Levels

We just saw that knowing the p-value allows us to accept or reject H_0 at level α.

However, the p-value is more informative than saying that we reject at level α.

It is the smallest level α at which we would still have rejected H_0.

It is also called the observed significance level.

Working with predefined α made it possible to choose the best level α test.
Best: Having highest probability of rejecting H_0 when H_1 is true.

This makes for nice and useful mathematical theory, but p-values should be the preferred way of judging and reporting test results for a specific test statistic. It may well be that a different test statistic $s_1(\mathbf{z})$ gives a smaller p-value than what the optimal test statistic $s_0(\mathbf{z})$ might give for a given $\mathbf{z}$.

This complicates finding optimal tests based on p-value behavior.

The Power Function

The probability of rejecting H_0 is denoted by β. It is a function of the distributional model F governing **Z**, i.e., $\beta = \beta(F)$. It is called the power function of the test.

When the hypothesis H_0 is composite and when $s(\mathbf{Z})$ has more than one possible distribution under H_0 one defines the highest probability of type I error as the significance level of the test. Hence $\alpha = \mathrm{maximum}\{\beta(F) : F \in H_0\}$.
α limits the type I error probability.

For various $F \in H_1$ the power function gives us the corresponding probabilities of type II error as $1 - \beta(F)$.

Note that some people denote the probability of type II error by $\beta = \beta(F)$.
Thus make sure what is meant by $\beta(F)$ when you read or write about it.

Samples and Populations

So far we have covered inference based on a randomization test. This relied heavily on our randomized assignment of flux X and flux Y to the 18 circuit boards.

Such inference can logically only say something about flux differences in the context of those 18 boards.

To generalize any conclusions to other boards would require some assumptions, judgement, and ultimately a step of faith.

Would the same conclusion have been reached for another set of 18 boards? What if one of the boards was an outlier board or something else was peculiar?

To be representative of the population of all boards we should view these 18 boards and their processing as a random sample from a conceptual population of such processed boards.

Conceptual Populations

Clearly the 18 boards happened to be available at the time of the experiment.

They could have been a random sample of all boards available at the time.
However, they also may have been taken sequentially in the order of production.
They certainly could not be a sample from future boards, yet to be produced.
They could not be a sample of boards already in use on aircrafts.

The randomized processing steps might give the appearance of random samples,
assuming that these steps are mostly responsible for response variations.

Thus we could regard the 9+9 SIR values as two random samples from two
very large or infinite conceptual populations of SIR values.
One sample of 9 boards from all boards/processes treated with flux X and
one sample of 9 boards from all boards/processes treated with flux Y.
A board can only be treated with flux X or Y (not both at the same time)
$\Rightarrow$ further conceptualization.

Population Distributions and Densities

Such infinite populations of Z-values are conveniently described by densities $f(z)$, with the properties $f(z) \geq 0$ and $\int_{-\infty}^{\infty} f(z)dz = 1$.

The probability of observing a randomly chosen element Z with $Z \leq x$ is

$$F(x) = P(Z \leq x) = \int_{-\infty}^{x} f(z)dz = \int_{-\infty}^{x} f(t)dt$$

z & t are just dummy variables. Avoid using x as dummy integration variable.

$F(x)$ as a function of x is also called the cumulative distribution function (CDF) of the random variable Z. $F(x) \nearrow$ from 0 to 1 as x goes from $-\infty$ to ∞.

For discrete populations with a finite or countably infinite number of distinct possible values z, we replace $f(z)$ by the probability mass function $p(z) = P(Z = z) \geq 0$ and write

$$F(x) = P(Z \leq x) = \sum_{z \leq x} p(z) \qquad \text{with} \qquad \sum_{z} p(z) = 1 \, .$$

Means, Expectations and Variances

The mean or expectation of Z or its population is defined by

$$\mu = \mu_Z = E(Z) = \int_{-\infty}^{\infty} z f(z) dz \quad \text{or} \quad \mu = E(Z) = \sum_z z p(z)$$

a probability weighted average of z values $=$ center of probability mass balance.

By extension, the mean or expectation of $g(Z)$ is defined by

$$E(g(Z)) = \int_{-\infty}^{\infty} g(z) f(z) dz \quad \text{or} \quad E(g(Z)) = \sum_z g(z) p(z)$$

Using $g(z) = (z - \mu)^2$ the variance of Z is defined by

$$\sigma^2 = \mathrm{var}(Z) = E\left((Z-\mu)^2\right) = \int_{-\infty}^{\infty} (z-\mu)^2 f(z) dz \quad \text{or} \quad \sigma^2 = \sum_z (z-\mu)^2 p(z)$$

$\sigma = \sigma_Z = \sqrt{\mathrm{var}(Z)}$ is called the standard deviation of Z or its population.

It is a measure of distribution spread.

p-Quantiles z_p

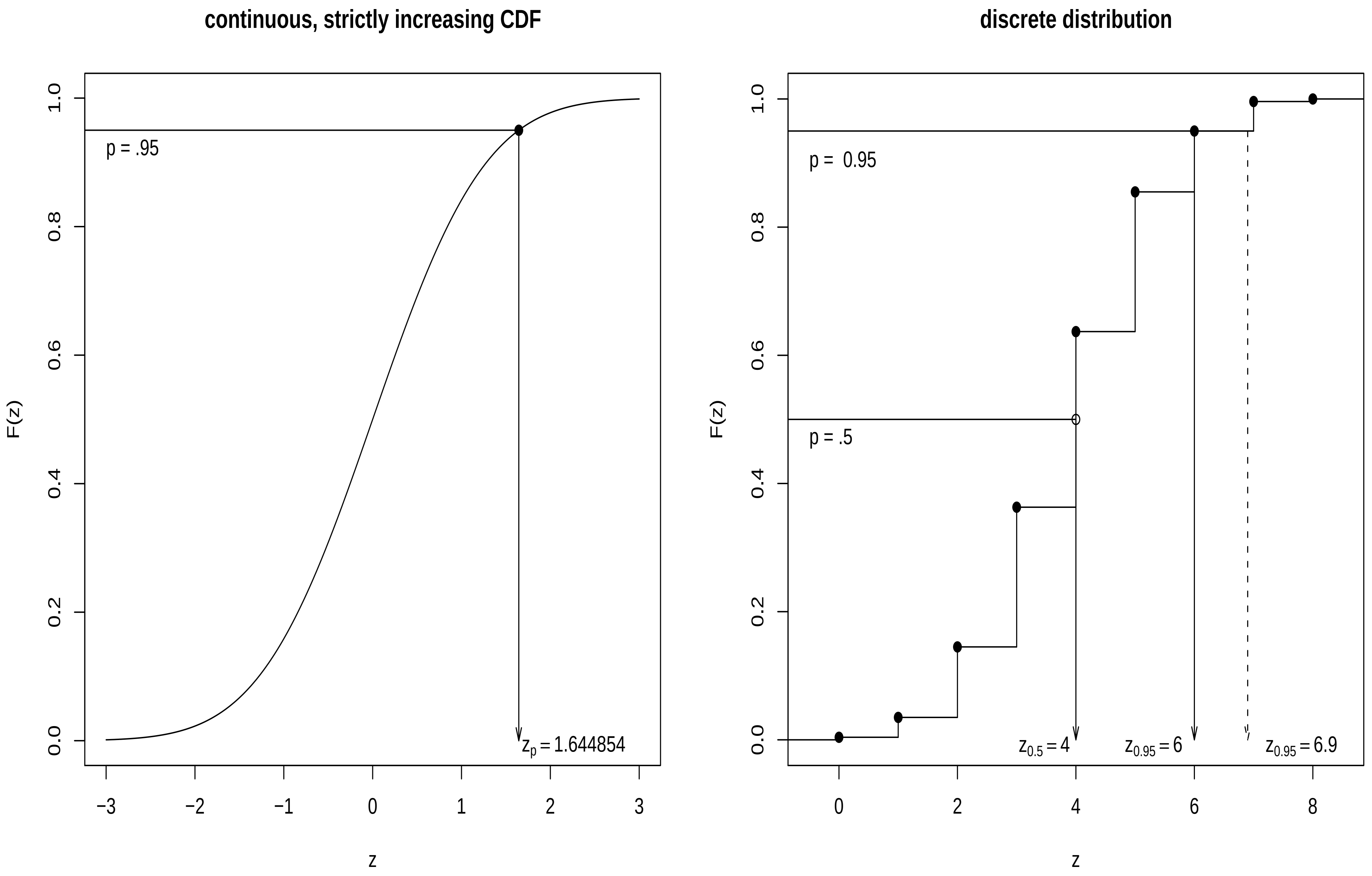
The p-quantile z_p of a distribution is defined as the lowest point z with $F(z) \geq p$.

When the distribution of Z is continuous and strictly increasing, z_p is defined by $F(z_p) = p$. This is the case typically encountered in this course.

If there is a whole interval $[z_1, z_2]$ over which we have $F(z) = p$ for $z \in [z_1, z_2]$ then any point in that interval could qualify as p-quantile, although the lowest point (z_1) is customarily chosen.
For $p = .5$ one may prefer the midpoint of such an interval.

If Z has a discrete distribution, i.e., $F(z)$ has jumps, then z_p could coincide with one of the jump points and we could have $F(z_p) > p$.

p-Quantile Illustrations



continuous, strictly increasing CDF

discrete distribution

Multivariate Densities or Populations

$f(z_1,\ldots,z_n)$ is a multivariate density if it has the following properties:

$f(z_1,\ldots,z_n) \geq 0$ for all $z_1,\ldots,z_n$ and $\int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(z_1,\ldots,z_n)\, dz_1\ldots dz_n = 1\,.$

It describes the behavior of the infinite population of such n-tuples $(z_1,\ldots,z_n)$.

A random element $(Z_1,\ldots,Z_n)$ drawn from such a population is a random vector.

We say that $Z_1,\ldots,Z_n$ in such a random vector are (statistically) independent when the following property holds:

$$f(z_1,\ldots,z_n) = f_1(z_1)\times\cdots\times f_n(z_n)$$

Here $f_i(z_i)$ is the marginal density of Z_i. It is obtainable from the multivariate density by integrating out all other variables, e.g.,

$$f_2(z_2) = \int_{-\infty}^{\infty}\ldots\int_{-\infty}^{\infty} f(z_1,z_2,z_3,\ldots,z_n)\, dz_1 dz_3\ldots dz_n\,.$$

$E(g(Z_1,\ldots,Z_n))$ and $\text{cov}(Z_1,Z_2)$

In analogy to the univariate case we define

$$E(g(Z_1,\ldots,Z_n) = \int_{-\infty}^{\infty}\cdots\int_{-\infty}^{\infty} g(z_1,\ldots,z_n)\, f(z_1,\ldots,z_n)\, dz_1\ldots dz_n$$

In particular, using $g(z_1,z_2,\ldots,z_n) = z_1\cdot z_2$,

$$E(Z_1Z_2) = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_1\cdot z_2\, f(z_1,z_2)\, dz_1 dz_2$$

For independent Z_1 and Z_2 we have

$$\begin{aligned} E(Z_1Z_2) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_1\cdot z_2\, f(z_1,z_2)\, dz_1 dz_2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty} z_1\cdot z_2\, f_1(z_1) f_2(z_2)\, dz_1 dz_2 \\ &= \int_{-\infty}^{\infty} z_1\, f_1(z_1)\, dz_1 \cdot \int_{-\infty}^{\infty} z_2\, f_2(z_2)\, dz_2 = E(Z_1)E(Z_2) \end{aligned}$$

Define the covariance of Z_1 and Z_2 as

$$\text{cov}(Z_1,Z_2) = E\left[(Z_1 - E(Z_1))(Z_2 - E(Z_2))\right] = E(Z_1Z_2) - E(Z_1)E(Z_2)$$

Note that independent Z_1 and Z_2 $\quad\Longrightarrow\quad \text{cov}(Z_1,Z_2) = 0,\quad$ but not $\quad\Longleftarrow$

Also note $\text{cov}(Z_1,Z_1) = E(Z_1^2) - [E(Z_1)]^2 = \text{var}(Z_1)$.

More on Independence

Conventionally, first definitions of independence start with

$P(Z_1 \in A_1, \ldots, Z_n \in A_n) = P(Z_1 \in A_1) \times \ldots \times P(Z_n \in A_n)$ for given sets $A_1, \ldots, A_n$,

but we get this from the density factorization as

$$P(Z_1 \in A_1, \ldots, Z_n \in A_n) = \int_{A_1} \cdots \int_{A_n} f(z_1, \ldots, z_n) dz_1 \ldots dz_n$$

$$= \int_{A_1} \cdots \int_{A_n} f_1(z_1) \times \cdots \times f_n(z_n) dz_1 \ldots dz_n = \int_{A_1} f_1(z_1) dz_1 \times \cdots \times \int_{A_n} f_n(z_n) dz_n$$

$$= P(Z_1 \in A_1) \times \cdots \times P(Z_n \in A_n)$$

By using $A_i = [x_i, x_i + h]$, $i = 1, \ldots, n$ we have from the conventional definition

$$\begin{aligned} P(Z_1 \in A_1, \ldots, Z_n \in A_n) &= \int_{A_1} \cdots \int_{A_n} f(z_1, \ldots, z_n) dz_1 \ldots dz_n \approx f(x_1, \ldots, x_n) h^n \\ P(Z_1 \in A_1) \times \cdots \times P(Z_n \in A_n) &= \int_{A_1} f_1(z_1) dz_1 \times \cdots \times \int_{A_n} f_n(z_n) dz_n \\ &\approx h^n f_1(x_1) \times \cdots \times f_n(x_n) \end{aligned}$$

$\approx$'s $\longrightarrow =$ at continuity points of f $\Longrightarrow$ $f(x_1, \ldots, x_n) = f_1(x_1) \times \cdots \times f_n(x_n)$.

Random Sample

When drawing repeatedly values $Z_1,\ldots,Z_n$ from a common infinite population with density $f(z)$ we get a multivariate random vector $(Z_1,\ldots,Z_n)$.

If the drawings are physically unrelated or “independent,” we may consider $Z_1,\ldots,Z_n$ as statistically independent, i.e., the random vector has density

$$h(z_1,\ldots,z_n) = f(z_1)\times\cdots\times f(z_n) \quad \text{note} \quad f_1 = \ldots = f_n = f\ .$$

$Z_1,\ldots,Z_n$ is then also referred to as a random sample from f.

We also express this as $Z_1,\ldots,Z_n \overset{\text{i.i.d.}}{\sim} f$.

Here i.i.d. = independent and identically distributed.

Rules of Expectations & Variances (Review)

For any set of random variables $X_1, \ldots, X_n$ and constants $a_0, a_1, \ldots, a_n$ we have

$$E\left(a_0 + a_1 \times X_1 + \ldots + a_n \times X_n\right) = a_0 + a_1 \times E(X_1) + \ldots + a_n \times E(X_n)$$

provided the expectations $E(X_1), \ldots, E(X_n)$ exist and are finite.

This holds whether $X_1, \ldots, X_n$ are independent or not.

For any set of independent random variables $X_1, \ldots, X_n$ and constants $a_0, a_1, \ldots, a_n$ we have

$$\mathrm{var}\left(a_0 + a_1 \times X_1 + \ldots + a_n \times X_n\right) = a_1^2 \times \mathrm{var}(X_1) + \ldots + a_n^2 \times \mathrm{var}(X_n)$$

provided the variances $\mathrm{var}(X_1), \ldots, \mathrm{var}(X_n)$ exist and are finite. $\mathrm{var}(a_0) = 0$.

This is also true under the weaker (than independence) condition $\mathrm{cov}(X_i, X_j) = E(X_i X_j) - E(X_i)E(X_j) = 0$ for $i \neq j$. In that case $X_1, \ldots, X_n$ are uncorrelated.

Rules for Averages

$$E\,(\bar{X}) = E\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n}E\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n}\sum_{i=1}^{n} E(X_i) = \frac{1}{n}\sum_{i=1}^{n}\mu_i = \bar{\mu}$$

whether $X_1,\ldots,X_n$ are independent or not.

If $\mu_1 = \ldots = \mu_n = \mu$ then $E(\bar{X}) = \mu$.

If $X_1,\ldots,X_n$ are independent (or uncorrelated) we also have

$$\mathrm{var}\,(\bar{X}) = \mathrm{var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\mathrm{var}\left(\sum_{i=1}^{n} X_i\right) = \frac{1}{n^2}\sum_{i=1}^{n}\mathrm{var}(X_i) = \frac{1}{n^2}\sum_{i=1}^{n}\sigma_i^2 = \frac{1}{n}\,\bar{\sigma}_n^2$$

where $\bar{\sigma}_n^2 = \frac{1}{n}\sum_{i=1}^{n}\sigma_i^2$. $\bar{\sigma}_n^2 = \sigma^2$ when $\sigma_1^2 = \ldots = \sigma_n^2 = \sigma^2$.

$\bar{\sigma}_n^2/n \searrow 0$ as $n \to \infty$, provided $\bar{\sigma}_n^2$ stays bounded, e.g., $\bar{\sigma}_n^2 = \sigma^2$.

A Normal Random Sample

$X_1,\ldots,X_n$ is called a normal random sample when the common density of the X_i is a normal density of the following form:

$$f(x) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right) \qquad \text{we also write} \quad X_i \sim \mathcal{N}(\mu,\sigma^2)\;.$$

This density or its associated population has mean μ and standard deviation σ.

When $\mu = 0$ and $\sigma = 1$, it is called the standard normal density

$$\varphi(x) = \frac{1}{\sqrt{2\pi}}\exp\left(-\frac{x^2}{2}\right) \qquad \text{with CDF} \qquad \Phi(x) = \int_{-\infty}^{x}\varphi(z)\,dz\;.$$

If $X \sim \mathcal{N}(\mu,\sigma^2)$ then $Z = (X-\mu)/\sigma \sim \mathcal{N}(0,1)$, the standard normal distribution

$\Rightarrow P(X \le x) = P((X-\mu)/\sigma \le (x-\mu)/\sigma) = \Phi((x-\mu)/\sigma)$.

The CLT & the Normal Population Model

The normal population model is motivated by the Central Limit Theorem (CLT).

This comes about because many physical or natural measured phenomena can be viewed as the addition of several independent source inputs or contributors.

$$Y = X_1 + \ldots + X_k \qquad \text{or} \qquad Y = a_0 + a_1 X_1 + \ldots + a_k X_k$$

for independent random variables $X_1, \ldots, X_k$ and constants $a_0, a_1, \ldots, a_k$.

Or in a 1-term Taylor expansion

$$\begin{aligned} Y = f(X_1, \ldots, X_k) &\approx f(\mu_1, \ldots, \mu_k) + \sum_{i=1}^{k} (X_i - \mu_i) \frac{\partial f(\mu_1, \ldots, \mu_k)}{\partial \mu_i} \\ &= a_0 + a_1 X_1 + \ldots + a_k X_k \end{aligned}$$

provided the linearization is sufficiently good, i.e., the deviations $X_i - \mu_i$ are small compared to the curvature of f (small σ_i).

Central Limit Theorem (CLT) I

- Suppose we randomly and independently draw random variables $X_1, \ldots, X_n$ from n possibly different populations with respective means $\mu_1, \ldots, \mu_n$ and standard deviations $\sigma_1, \ldots, \sigma_n$

- Suppose further that the following variance ratio property holds

$$\max_{i=1,\ldots,n} \left(\frac{\sigma_i^2}{\sigma_1^2 + \ldots + \sigma_n^2} \right) \to 0 \, , \quad \text{as} \quad n \to \infty$$

  i.e., none of the variances dominates among all variances, obviously the case when $\sigma_1 = \ldots = \sigma_n$ and $n \to \infty$.

- Then $Y = Y_n = X_1 + \ldots + X_n$ has an approximate normal distribution with mean and variance given by

$$\mu_Y = \mu_1 + \ldots + \mu_n \quad \text{and} \quad \sigma_Y^2 = \sigma_1^2 + \ldots + \sigma_n^2 \, .$$

Central Limit Theorem (CLT) II

The next few slides illustrate how well the CLT performs and when it falls short.

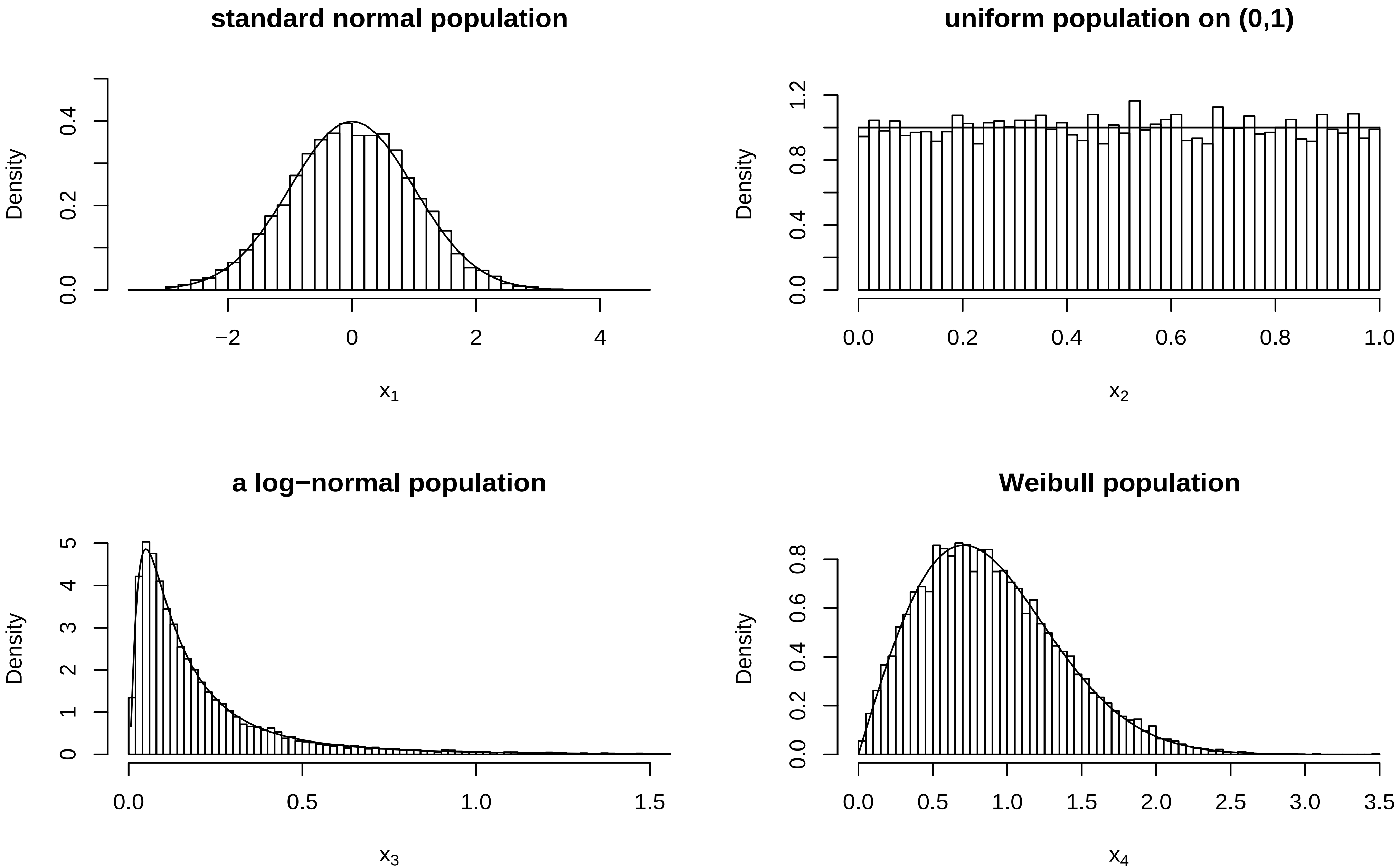
In the following CLT illustrations the superimposed normal density corresponds to $\mu = \mu_1 + \ldots + \mu_n$ and $\sigma^2 = \sigma_1^2 + \ldots + \sigma_n^2$.

First we show 4 (5) distributions from which the $X_1, \ldots, X_4$ are sampled, respectively. These distributions cover a wide spectrum of shapes.

We only sample 4, illustrating that $n = 4$ can be sufficiently close to ∞ in order for the CLT to become reasonably effective.

Note the slight deterioration in the normal appearance when we exchange the normal distribution with a skewed distribution, i.e. when generating $X_2 + \ldots + X_5$ instead of $X_1 + \ldots + X_4$. ($\Longrightarrow$ slide 30 (VI))

Central Limit Theorem (CLT)   III



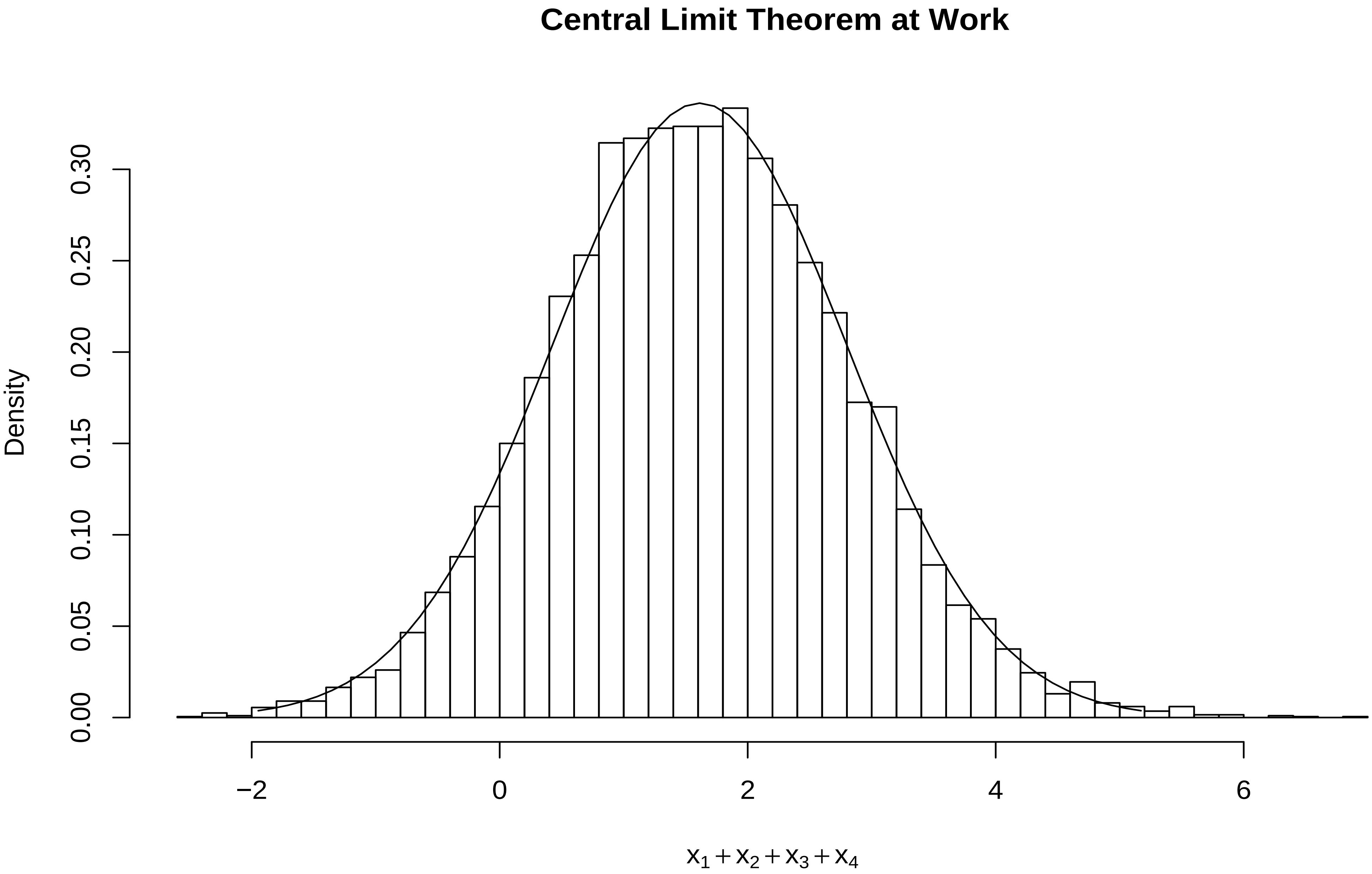
standard normal population

uniform population on (0,1)

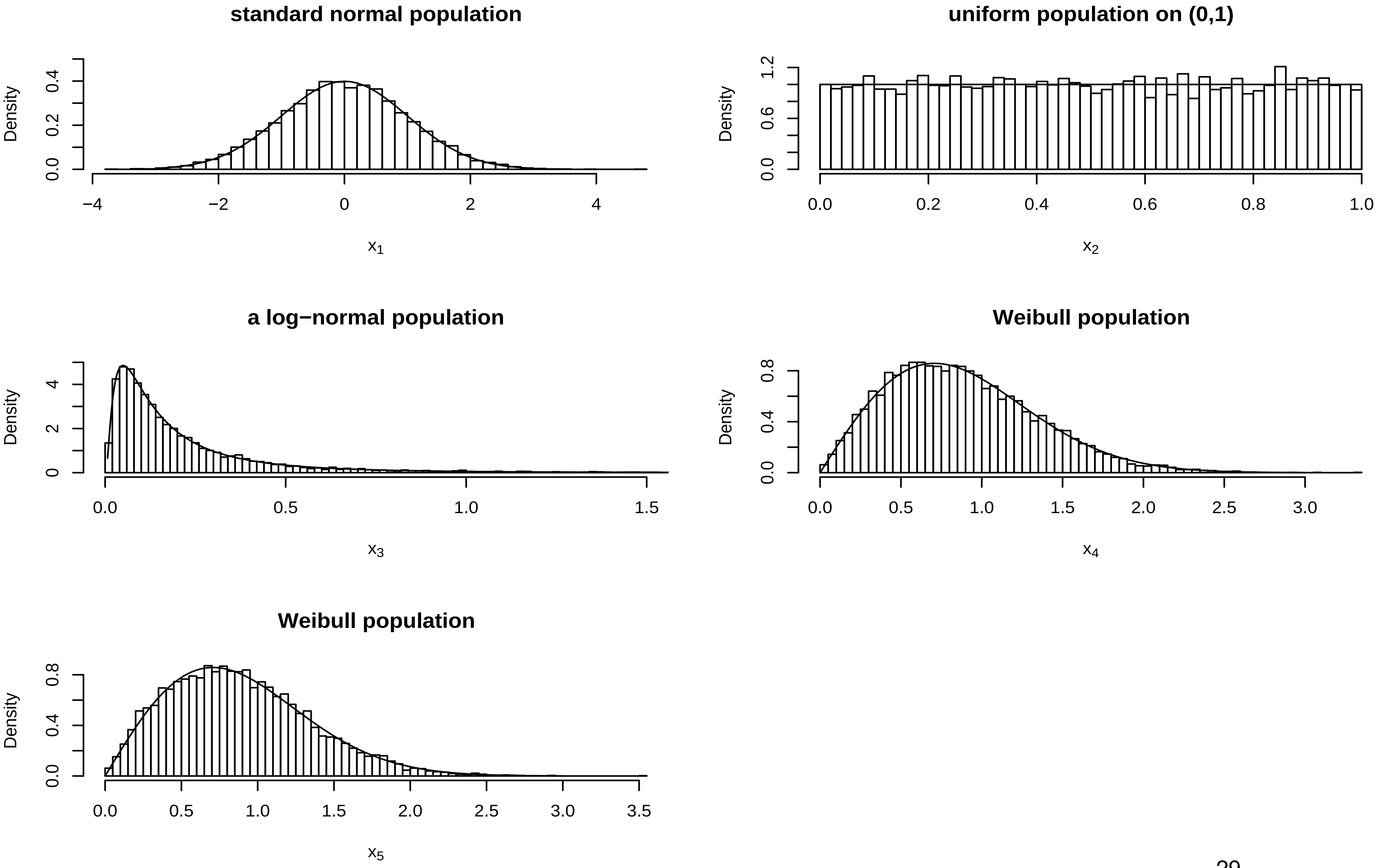
a log−normal population

Weibull population

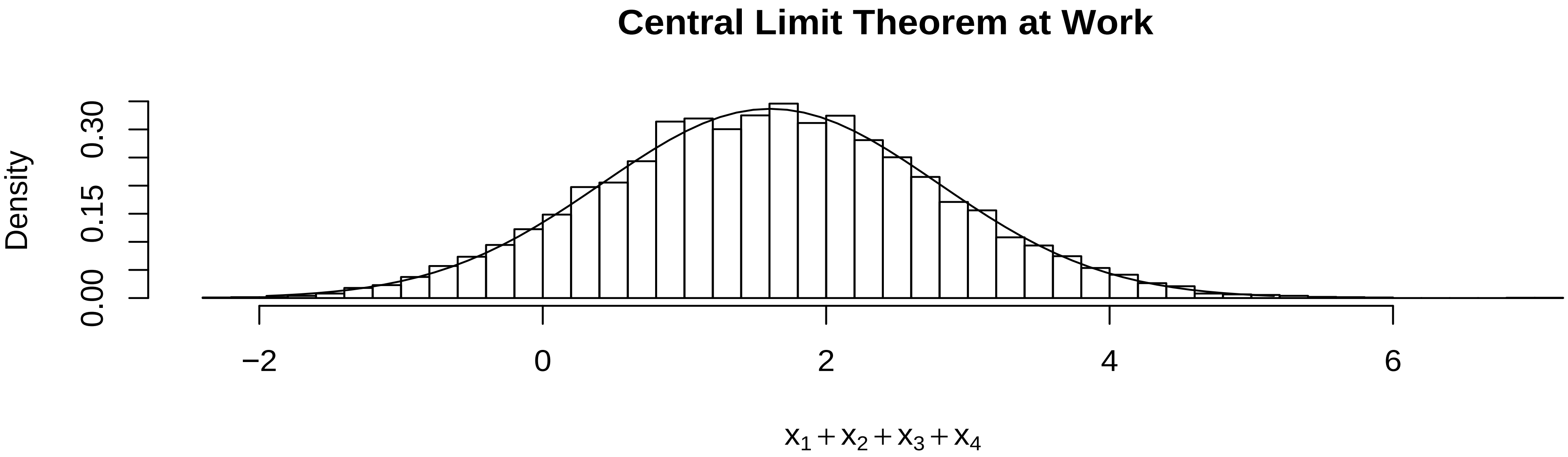
Central Limit Theorem (CLT) IV

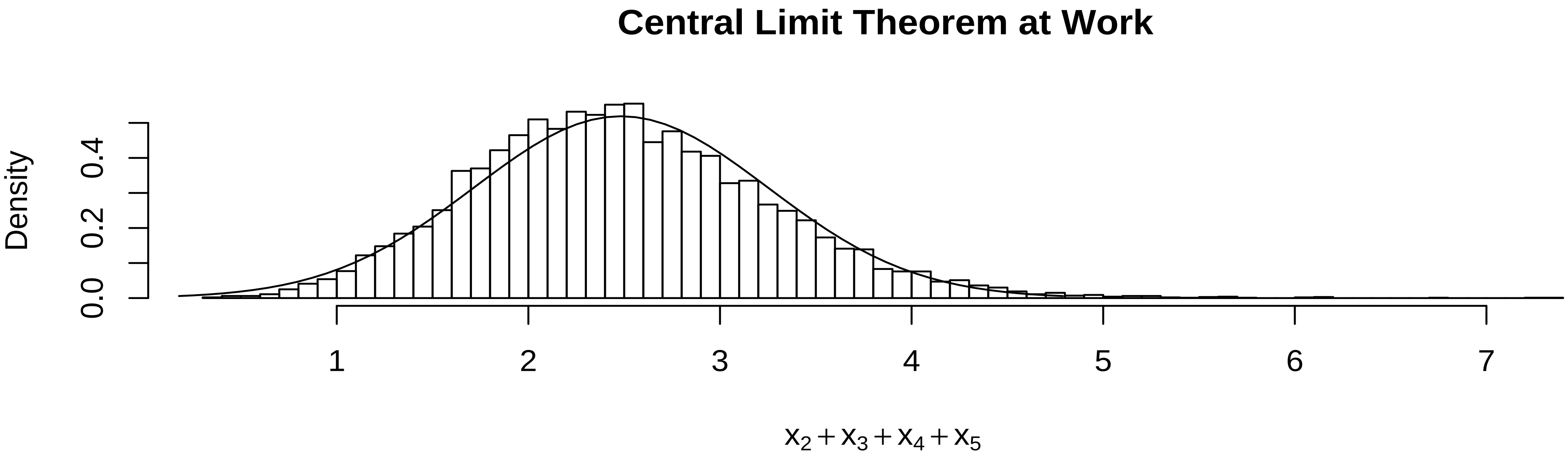


Central Limit Theorem (CLT) V

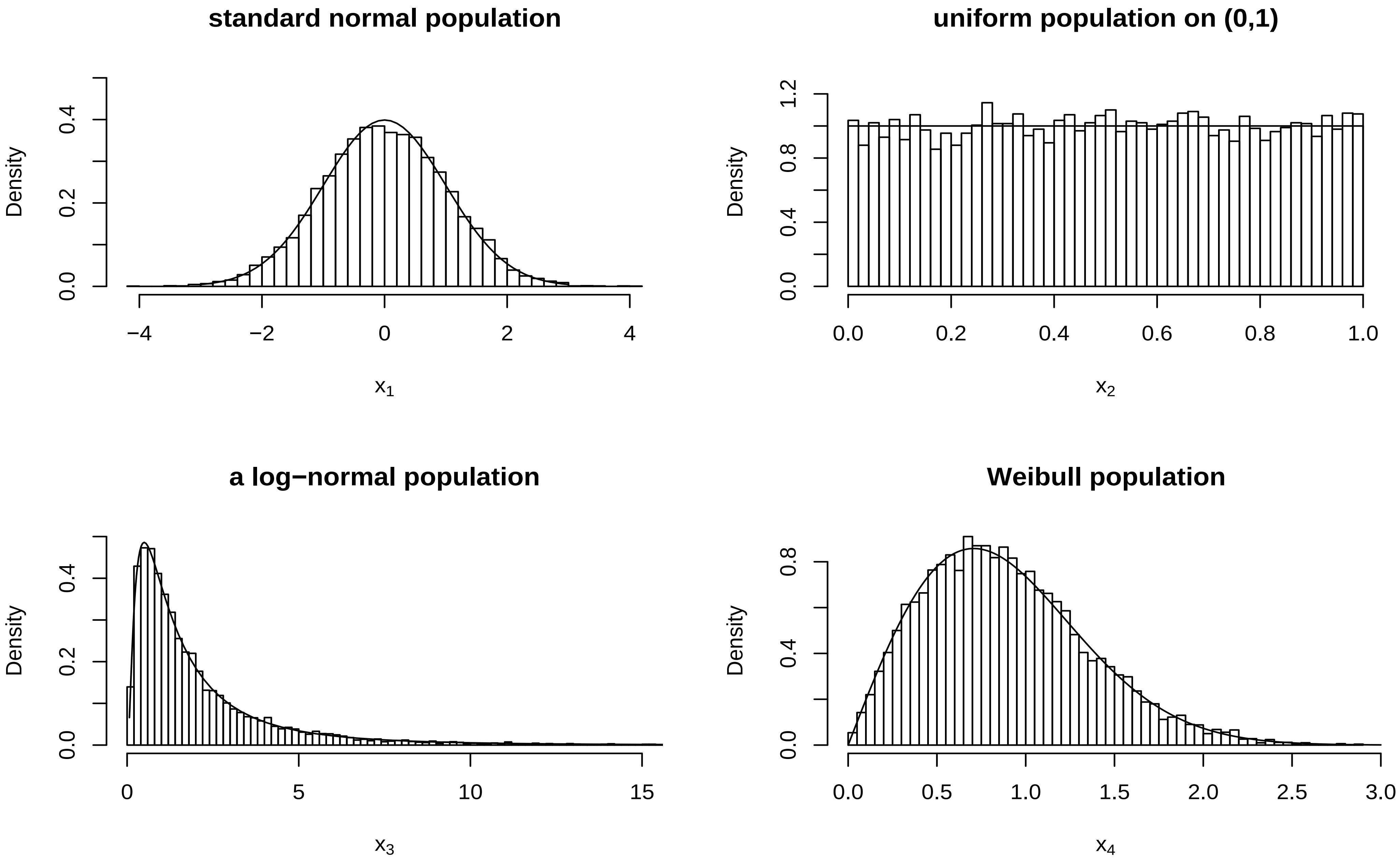


Central Limit Theorem (CLT) VI





Central Limit Theorem (CLT) VII



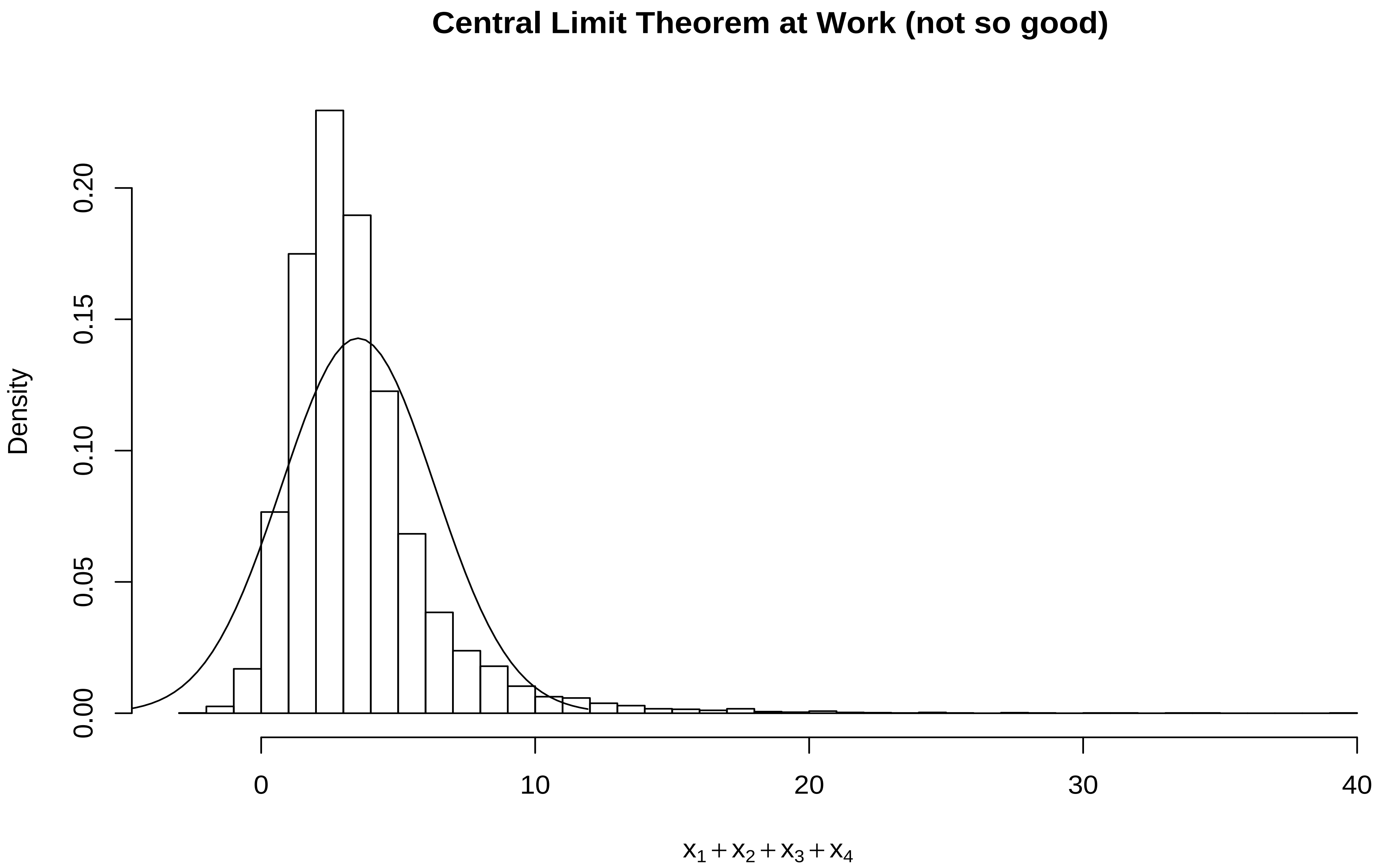
standard normal population

uniform population on (0,1)

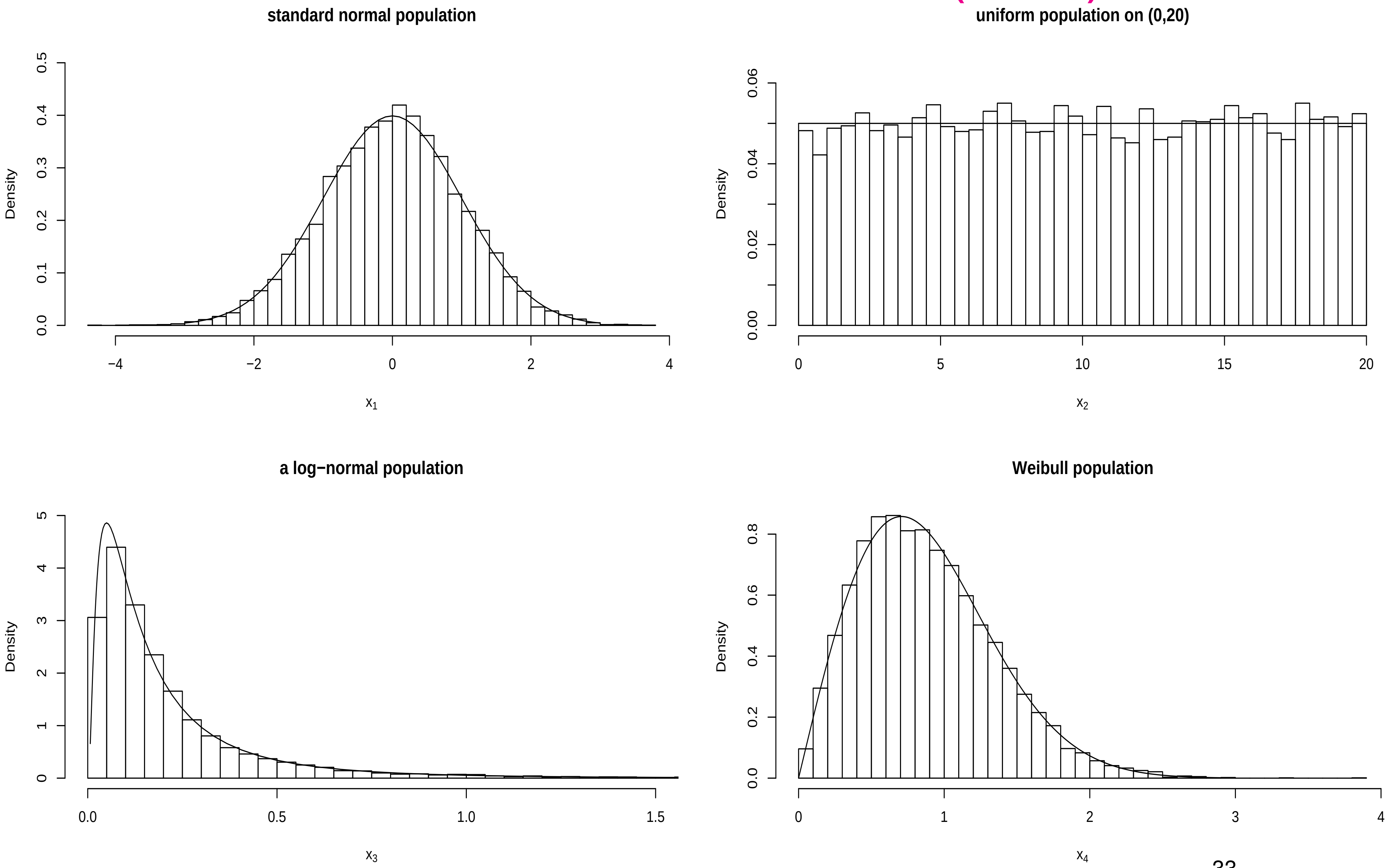
a log−normal population

Weibull population

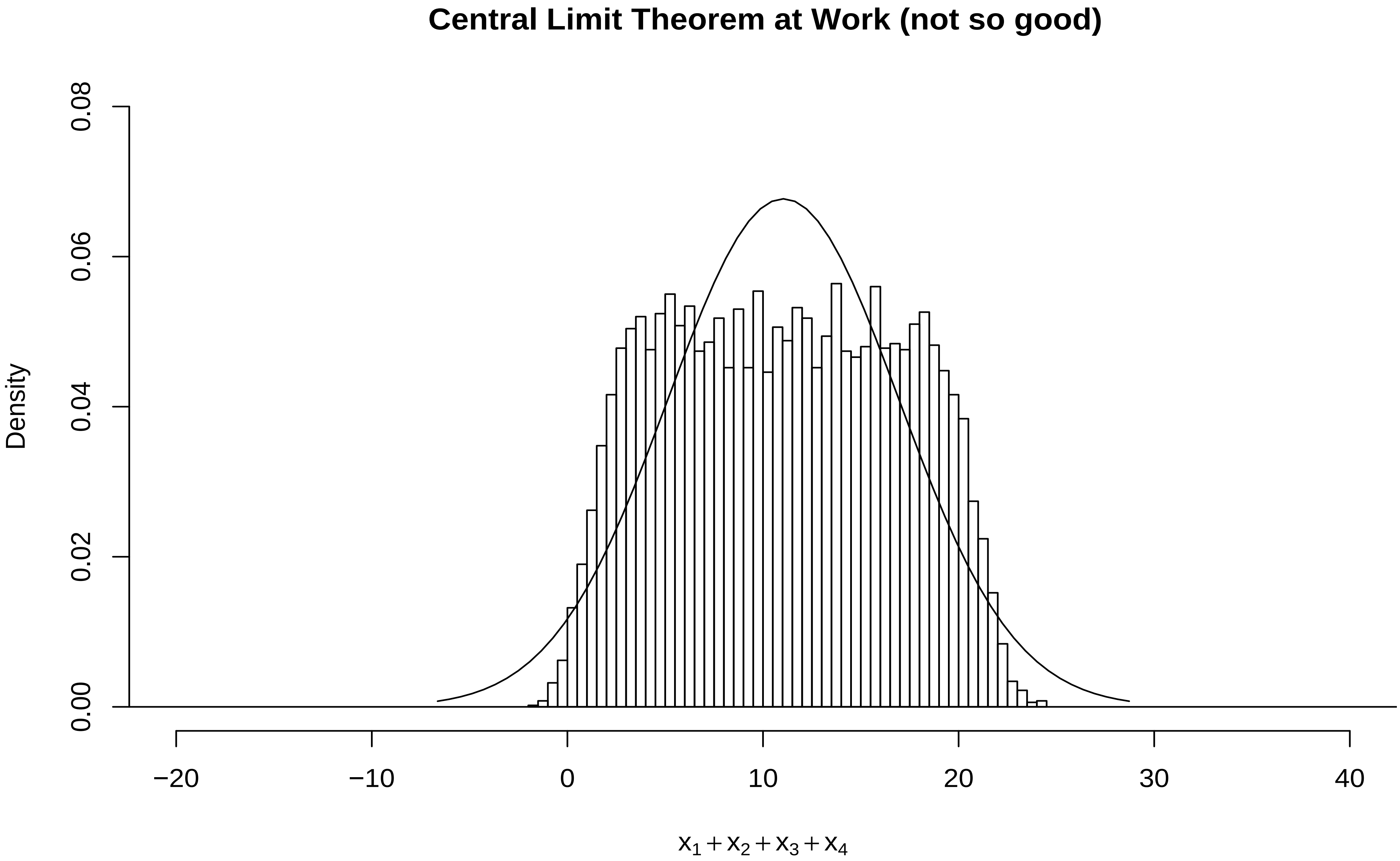
Central Limit Theorem (CLT) VIII



Central Limit Theorem (CLT) IX



Central Limit Theorem (CLT)   X



Central Limit Theorem at Work (not so good)

Central Limit Theorem (CLT) XI

The last 4 slides illustrate the result when the condition none of the variances dominates is violated.

First we scaled up the log-normal distribution by a factor of 10 and the resulting distribution of $X_1 + \ldots + X_4$ looks skewed to the right and looks in shape very much like the dominating log-normal variation source.

In the second such example we instead scaled up the uniform distribution by a factor of 20. The resulting distribution of $X_1 + \ldots + X_4$ looks almost like a uniform distribution with somewhat smoothed shoulders.

In both cases the distribution with the dominating variability imprints its character on the resulting distribution of $X_1 + \ldots + X_4$.

Central Limit Theorem (CLT) XII

What would happen if instead we increased the spread in X_1 by a factor of 10?

Central Limit Theorem (CLT) XIII

What would happen if instead we increased the spread in X_1 by a factor of 10?

The distribution of $X_1 + \ldots + X_4$ would look normal,

like the dominating X_1 distribution.

Derived Distributions from Normal Model

Other than working with randomization reference distributions we otherwise generally assume normal distributions as the sources of our data

Thus it is worthwhile to characterize some sampling distributions that are derived from the normal distribution. They will play a significant role later on.

The chi-square distribution, the Student t-distribution, and the F-distribution.

These distributions come about as sampling distributions of certain test statistics based on normal random samples.

Much of what is covered in this course could also be dealt with in the context of other sampling population models. We will not get into that.

Properties of Normal Random Variables

Assume that $X_1, \ldots, X_n$ are independent normal random variables with respective means and variances given by: $\mu_1, \ldots, \mu_n$ and $\sigma_1^2, \ldots, \sigma_n^2$. Then

$Y = X_1 + \ldots + X_n \sim \mathcal{N}(\mu_1 + \ldots + \mu_n, \sigma_1^2 + \ldots + \sigma_n^2)$ Geometric proof in Appendix A

Here $\sim$ means " exactly distributed as"

If $X \sim \mathcal{N}(\mu, \sigma^2)$ then $a + bX \sim \mathcal{N}(a + b\mu, b^2\sigma^2)$

$$\frac{X - \mu}{\sigma} \sim \mathcal{N}(0,1) \qquad \text{with} \quad b = 1/\sigma \quad \text{and} \quad a = -\mu/\sigma$$

Caution: Some people write $X \sim \mathcal{N}(\mu, \sigma)$ when others (and I) write $X \sim \mathcal{N}(\mu, \sigma^2)$.

For example, in R we have `dnorm(x,`μ`,`σ`)`, `pnorm(x,`μ`,`σ`)`,`qnorm(p,`μ`,`σ`)`,and `rnorm(n,`μ`,`σ`)`, which respectively give the density, CDF, quantile of $\mathcal{N}(\mu, \sigma^2)$ and random samples from $\mathcal{N}(\mu, \sigma^2)$.

The Chi-Square Distribution

When $Z_1,\ldots,Z_f \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$ we say that

$$C_f = \sum_{i=1}^{f} Z_i^2$$ has a chi-square distribution with f degrees of freedom

Memorize this definition! We also write $C_f \sim \chi^2_f$. The density of C_f is

$$h(x) = \frac{1}{2^{f/2}\Gamma(n/2)} x^{f/2-1}\exp(-x/2) \quad \text{for } x > 0$$ not to memorize

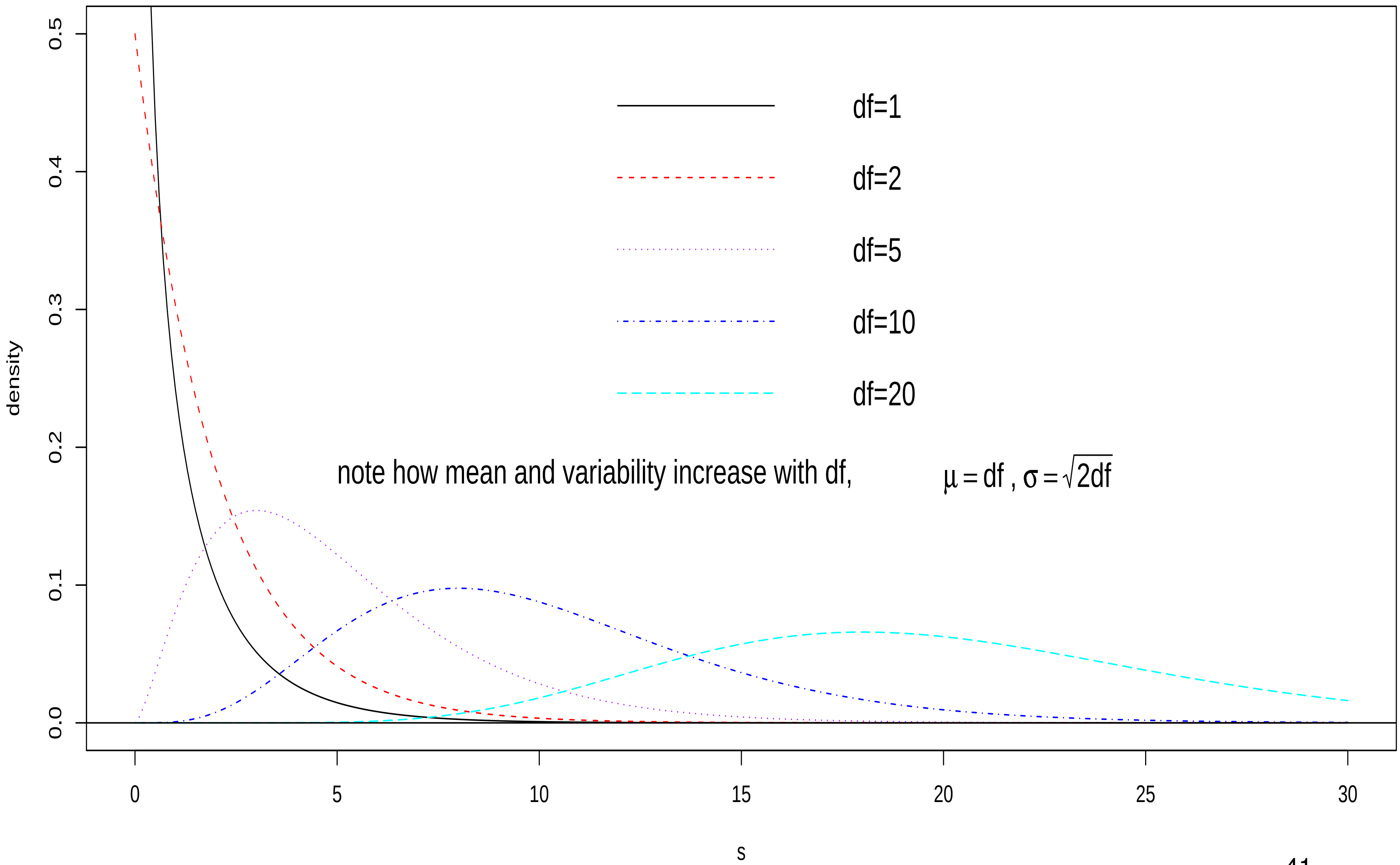
with mean $= f$ and variance $= 2f$, worth memorizing.

Density, CDF, quantiles of, and random samples from the chi-square distribution can be obtained in R via: `dchisq(x,f)`,`pchisq(x,f)`,`qchisq(p,f)`,`rchisq(N,f)`.

If $C_{f_1} \sim \chi^2_{f_1}$ and $C_{f_2} \sim \chi^2_{f_2}$ are independent then $C_{f_1} + C_{f_2} \sim \chi^2_{f_1+f_2}$

since $Z_1^2 + \ldots + Z_{f_1}^2 + \tilde{Z}_1^2 + \ldots + \tilde{Z}_{f_2}^2 \sim \chi^2_{f_1+f_2}$ with Z_i, $\tilde{Z}_j$ independent $\sim \mathcal{N}(0,1)$.

χ^2 Densities



The Noncentral χ^2_f Distribution

Suppose $X_1 \sim \mathcal{N}(d_1, 1), \ldots, X_f \sim \mathcal{N}(d_f, 1)$ are independent. Then we say that

$$C = X_1^2 + \ldots + X_f^2 \sim \chi^2_{f,\lambda}$$

has a noncentral χ^2 distribution with f degrees of freedom

and noncentrality parameter $\lambda = \sum_{i=1}^{f} d_i^2$. (memorize definition!)

$E(C) = f + \lambda$ and $\mathrm{var}(C) = 2f + 4\lambda$.

$(d_1, \ldots, d_f) = (0, \ldots, 0) \Longrightarrow$ previously defined (central) χ^2_f distribution.

The distribution of C depends on $(d_1, \ldots, d_f)$ only through $\lambda = \sum_{i=1}^{f} d_i^2$!

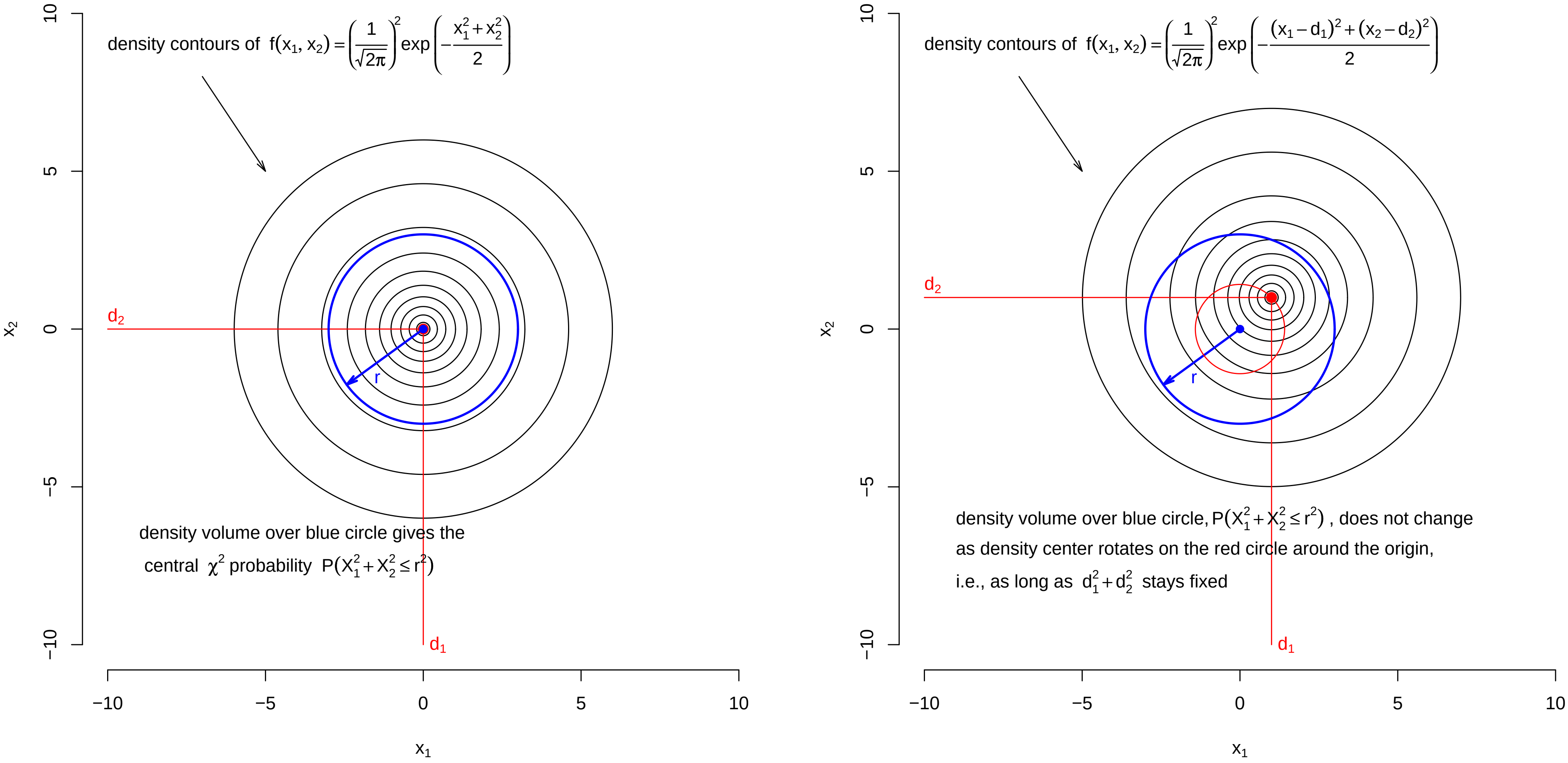
See geometric explanation on next slide.

What does R give us?

For the noncentral χ^2 we have: `dchisq(x, df, ncp=0),`

`pchisq(q, df, ncp=0), qchisq(p, df, ncp=0), rchisq(n, df, ncp=0)`

Dependence on $d_1^2 + d_2^2 = \lambda$ Only



density contours of $f(x_1, x_2) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left(-\frac{x_1^2 + x_2^2}{2}\right)$

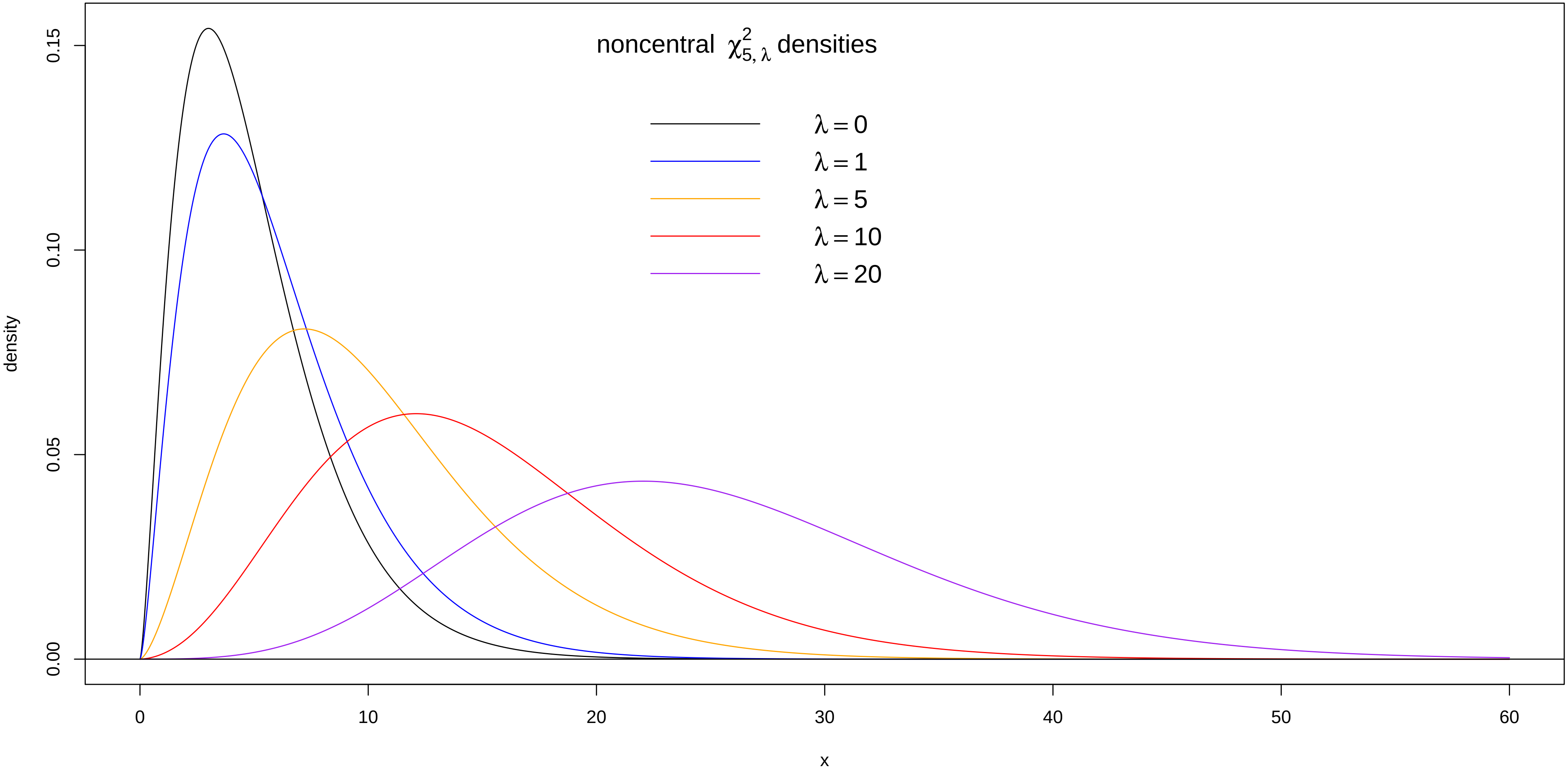
density volume over blue circle gives the central χ^2 probability $P(X_1^2 + X_2^2 \le r^2)$

density contours of $f(x_1, x_2) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp\left(-\frac{(x_1 - d_1)^2 + (x_2 - d_2)^2}{2}\right)$

density volume over blue circle, $P(X_1^2 + X_2^2 \le r^2)$, does not change as density center rotates on the red circle around the origin, i.e., as long as $d_1^2 + d_2^2$ stays fixed

This works because of the rotational invariance of density & the blue circle region.

Noncentral χ^2 Densities



The Student t-Distribution

When $Z \sim \mathcal{N}(0,1)$ is independent of $C_f \sim \chi^2_f$ we say that

$$t = \frac{Z}{\sqrt{C_f/f}}$$ memorize this definition!

has a Student t-distribution with f degrees of freedom. We also write $t \sim t_f$.
Its density is

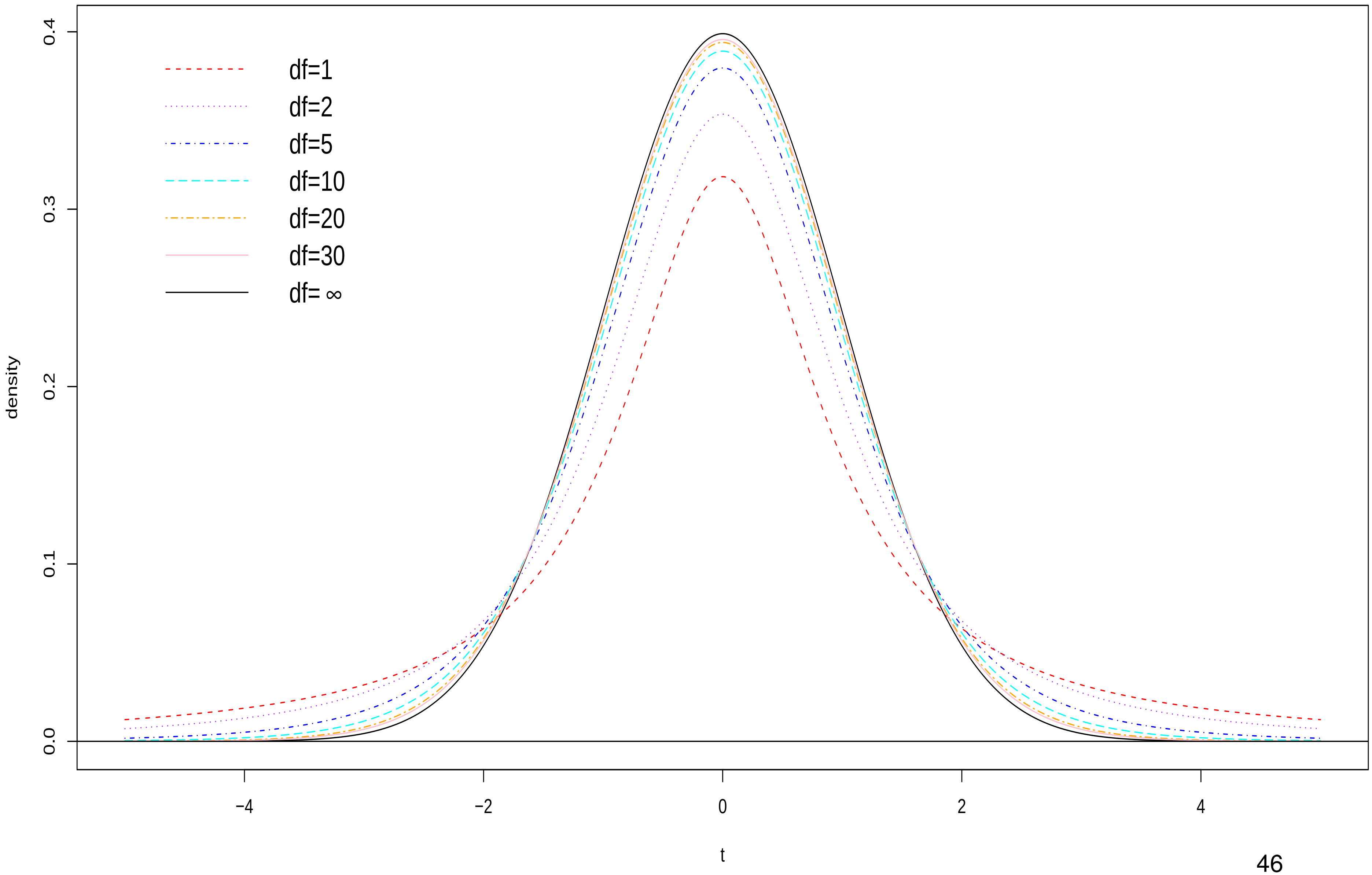
$$g_f(x) = \frac{\Gamma((f+1)/2)}{\sqrt{f\pi}\,\Gamma(f/2)} \left[x^2/f+1\right]^{-(f+1)/2} \quad \text{for } -\infty < x < \infty$$ not to memorize

It has mean 0 (for $f > 1$) and variance $f/(f-2)$ if $f > 2$.

$g_f(x) \longrightarrow \varphi(x)$ (standard normal density) as $f \longrightarrow \infty$. This follows either directly from the density using $(1+x^2/f)^{-f/2} \longrightarrow \exp(-x^2/2)$, or the definition of t because $C_f/f \longrightarrow 1$ which follows from $E(C_f/f) = 1$ & $\mathrm{var}(C_f/f) = 2f/f^2 \longrightarrow 0$.

Density, CDF, quantiles of, and random samples from the Student t-distribution can be obtained in R via: `dt(x,f)`, `pt(x,f)`, `qt(p,f)`, and `rt(N,f)` respectively.

Densities of the Student t-Distribution



The Noncentral Student t-Distribution

When $X \sim \mathcal{N}(\delta, 1)$ is independent of $C_f \sim \chi^2_f$ we say that

$$t = \frac{X}{\sqrt{C_f/f}} \qquad \text{memorize this definition!}$$

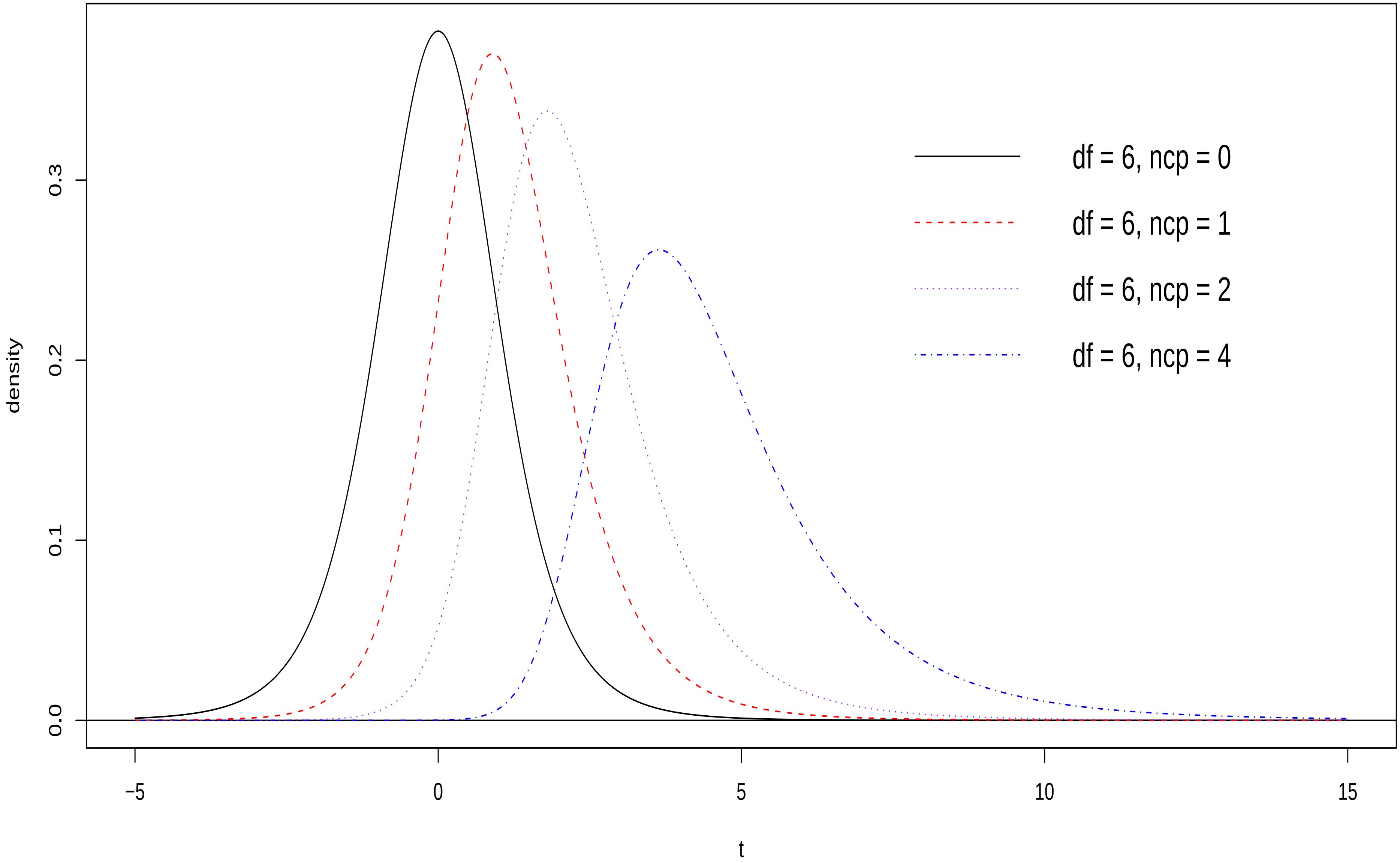
has a noncentral Student t-distribution with f degrees of freedom and noncentrality parameter `ncp` $=\delta$. We also write $t \sim t_{f,\delta}$.

R gives us:

`dt(x,f,ncp)`, `pt(q,f,ncp)`, `qt(p,f,ncp)` , and `rt(n,f,ncp)` respectively.

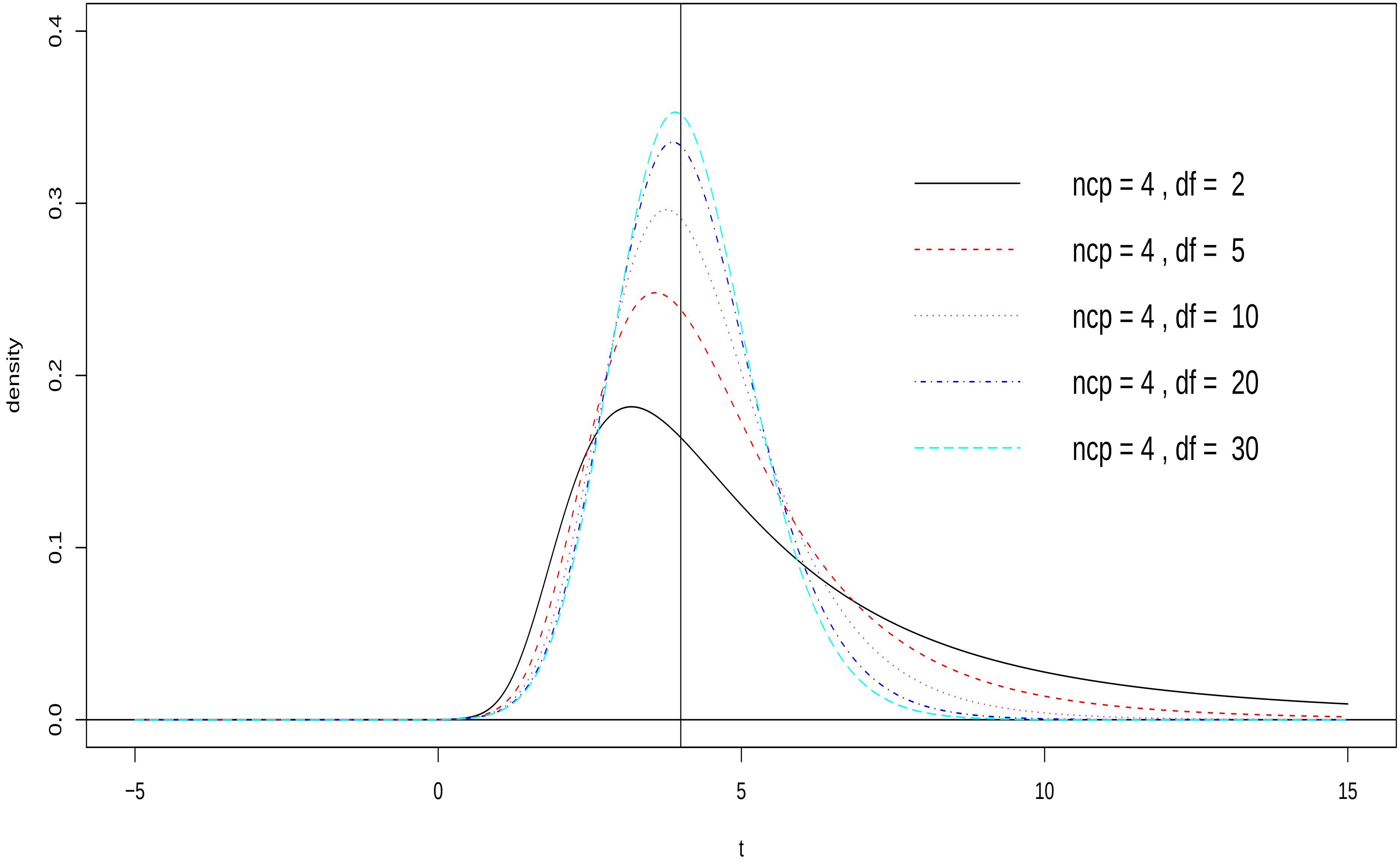
As before one can argue that this distribution converges to $\mathcal{N}(\delta, 1)$ as $f \to \infty$.

Densities of the Noncentral Student t-Distribution



These densities march to the left for negative `ncp`.

Densities of the Noncentral Student t-Distribution



The F-Distribution

When $C_{f_1} \sim \chi^2_{f_1}$ and $C_{f_2} \sim \chi^2_{f_2}$ are independent χ^2 random variables with f_1 and f_2 degrees of freedom, respectively, we say that

$$F = \frac{C_{f_1}/f_1}{C_{f_2}/f_2} \qquad \text{memorize this definition!}$$

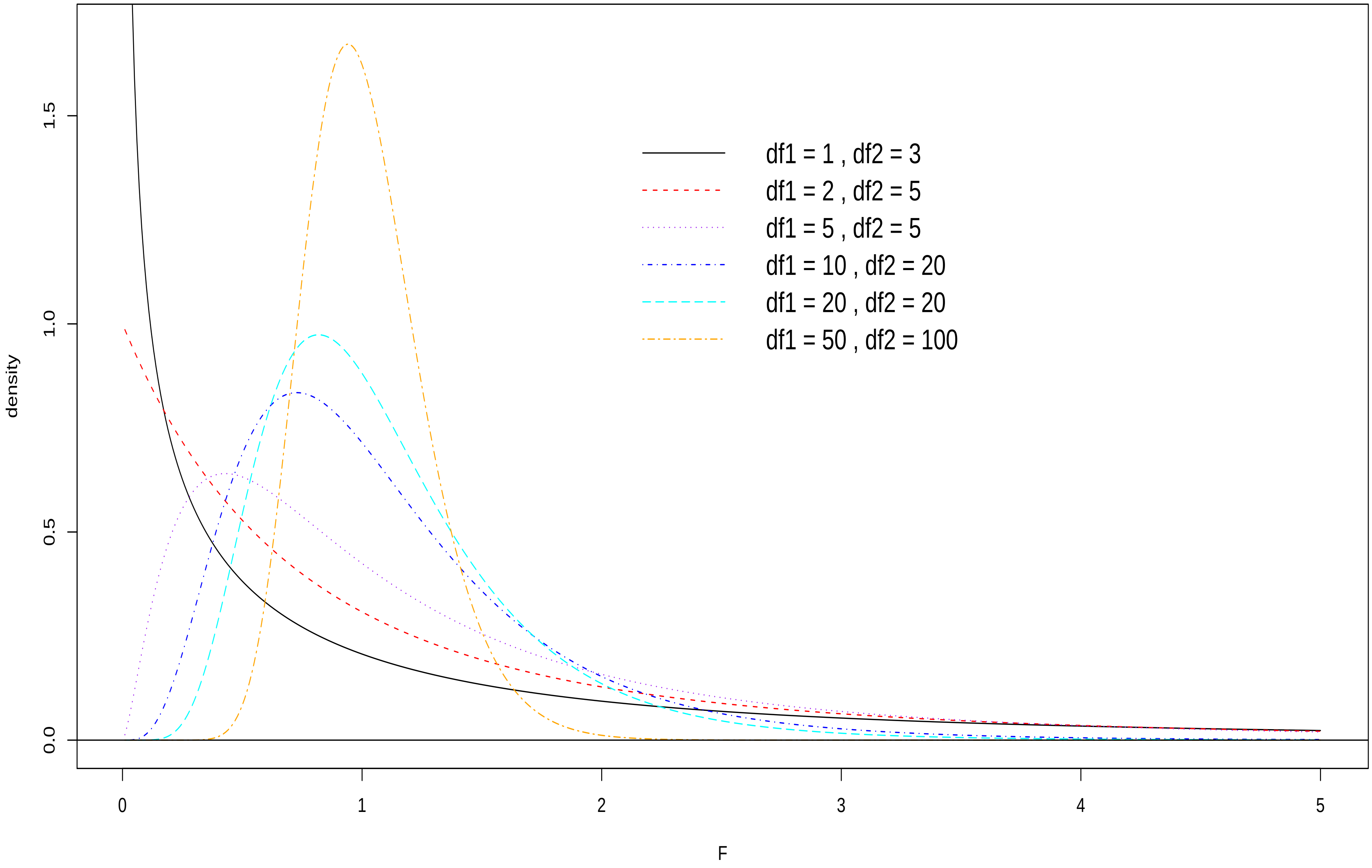
has an F distribution with f_1 and f_2 degrees of freedom. We also write $F \sim F_{f_1,f_2}$. Its density is

$$g(x) = \frac{\Gamma((f_1+f_2)/2)(f_1/f_2)^{f_1/2}\, x^{(f_1/2)-1}}{\Gamma(f_1/x)\Gamma(f_2/2)[(f_1/f_2)\,x+1]^{(f_1+f_2)/2}} \qquad \text{not to memorize}$$

Density, CDF, quantiles of,and random samples from the F_{f_1,f_2}-distribution can be obtained in R via: `df(x,f1,f2)`, `pf(x,f1,f2)`, `qf(p,f1,f2)`, `rf(N,f1,f2)`, respectively.

$t \sim t_f \implies t^2 \sim F_{1,f}$. Why? Also, $F \sim F_{f_1,f_2} \implies 1/F \sim F_{f_2,f_1}$. Why?

F Densities


df1 = 1 , df2 = 3
df1 = 2 , df2 = 5
df1 = 5 , df2 = 5
df1 = 10 , df2 = 20
df1 = 20 , df2 = 20
df1 = 50 , df2 = 100
density
0.0
0.5
1.0
1.5
F
0
1
2
3
4
5

The Noncentral F-Distribution

Let $C_1 \sim \chi_{f_1,\lambda}$ be a noncentral χ^2 random variable

and let $C_2 \sim \chi^2_{f_2}$ be a (central) χ^2 random variable which is independent of C_1,

then we say that

$$F = \frac{C_1/f_1}{C_2/f_2} \sim F_{f_1,f_2,\lambda}$$

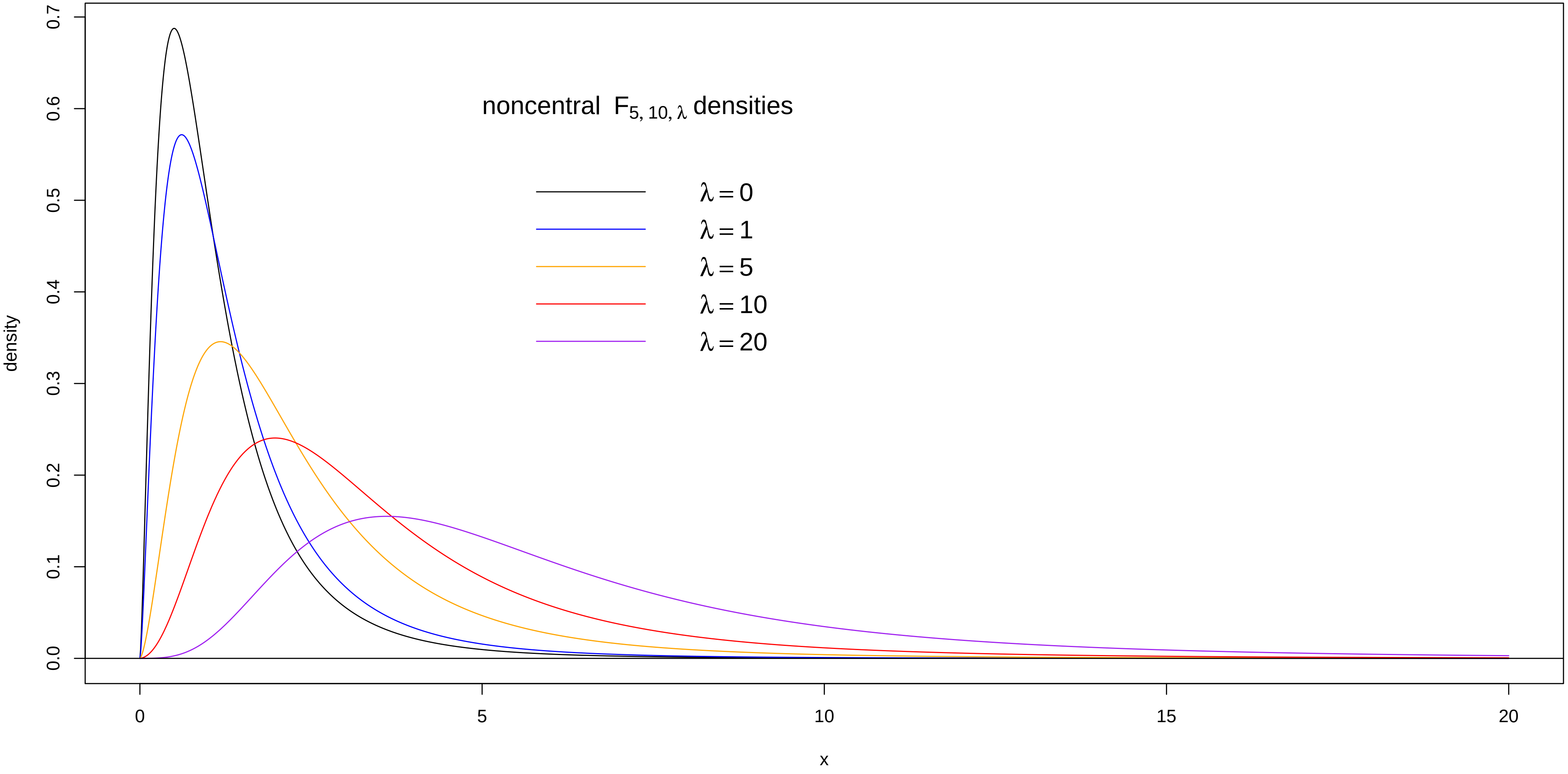
has a noncentral F-distribution with f_1 and f_2 degrees of freedom

and with noncentrality parameter λ. (memorize definition!)

What does R give us?

For the noncentral F we have: `df(x, df1, df2, ncp)`, `pf(q, df1, df2, ncp)`, `qf(p, df1, df2, ncp)`, `rf(n, df1, df2, ncp)`

Noncentral F Densities



Decomposition of the Sum of Squares (SS)

We illustrate here an early example of the SS-decomposition.

$$\begin{aligned}
\sum_{i=1}^{n} X_i^2 &= \sum_{i=1}^{n}(X_i-\bar{X}+\bar{X})^2 = \sum_{i=1}^{n}\left[(X_i-\bar{X})^2+2(X_i-\bar{X})\bar{X}+\bar{X}^2\right] \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2+2\sum_{i=1}^{n}(X_i-\bar{X})\bar{X}+\sum_{i=1}^{n}\bar{X}^2 \\
&= \sum_{i=1}^{n}(X_i-\bar{X})^2+\sum_{i=1}^{n}\bar{X}^2 = \sum_{i=1}^{n}(X_i-\bar{X})^2+n\bar{X}^2 \, .
\end{aligned}$$

since $\sum(X_i-\bar{X})=\sum X_i-n\bar{X}=\sum X_i-n\sum X_i/n=0$ i.e., the residuals sum to zero.

Such decompositions are the intrinsic and recurring theme in the Analysis of Variance (ANOVA) to be addressed at length later.

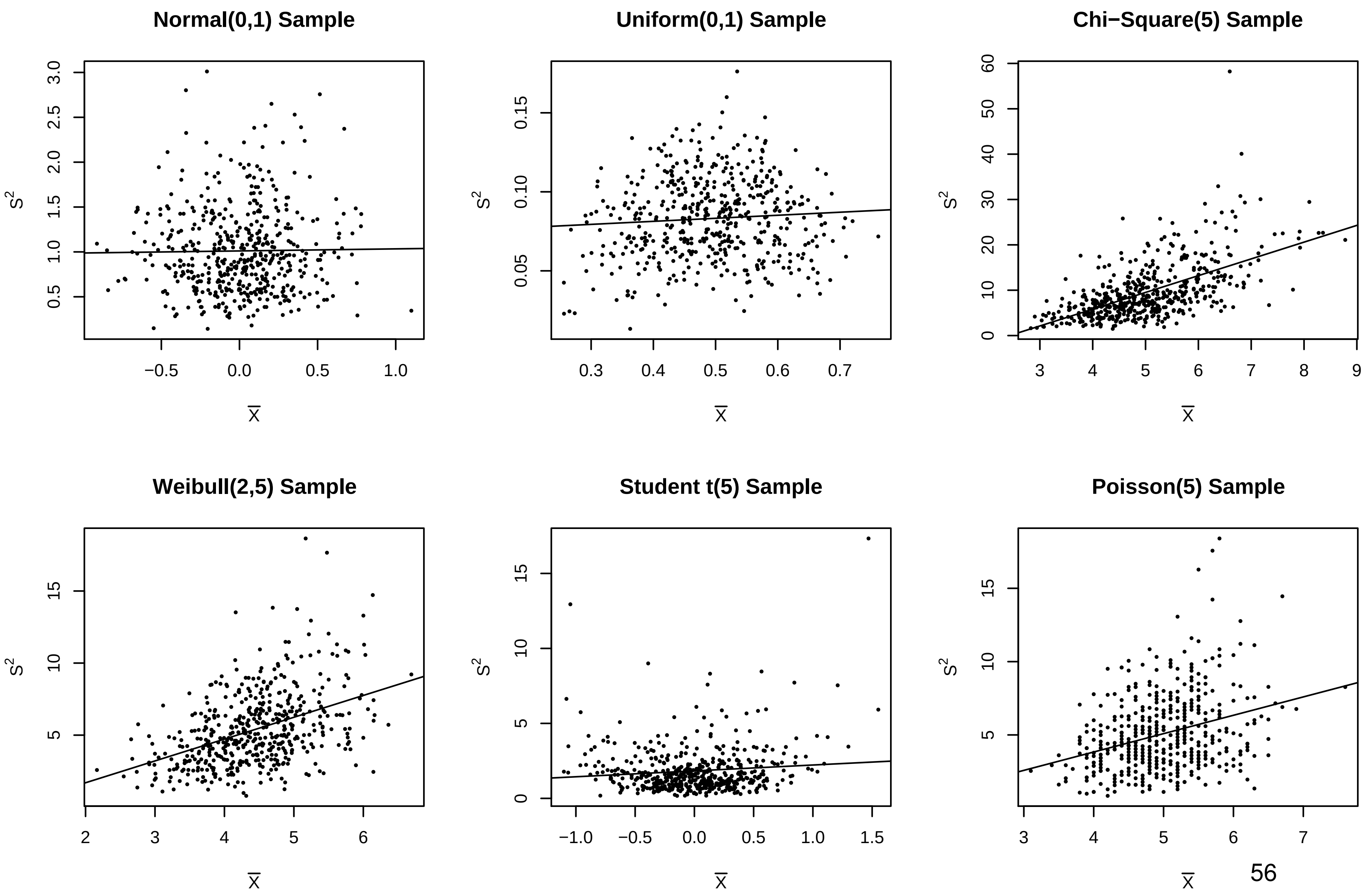
Sampling Distribution of $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$

When $X_1, \ldots, X_n$ are a random sample from $\mathcal{N}(\mu, \sigma^2)$ the joint distribution of $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ can be described as follows, as is shown in Appendix B.

- $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are statistically independent
- $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ or $(\bar{X} - \mu)/(\sigma/\sqrt{n}) = \sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$
- $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$ or $\sum_{i=1}^{n}(X_i - \bar{X})^2 \sim \sigma^2\chi^2_{n-1}$ (ambiguously)
  Only $n-1$ of the terms $X_i - \bar{X}$ can vary "independently," since $\sum(X_i - \bar{X}) = 0$.

The above independence seems perplexing, given that $\bar{X}$ also appears within the expression $\sum_{i=1}^{n}(X_i - \bar{X})^2$. It is a peculiar property of normal distribution samples. This independence will not occur for other sampled distributions.

$\bar{X}$ and S^2 Dependence (500 Samples of Size $n = 10$)



Comments

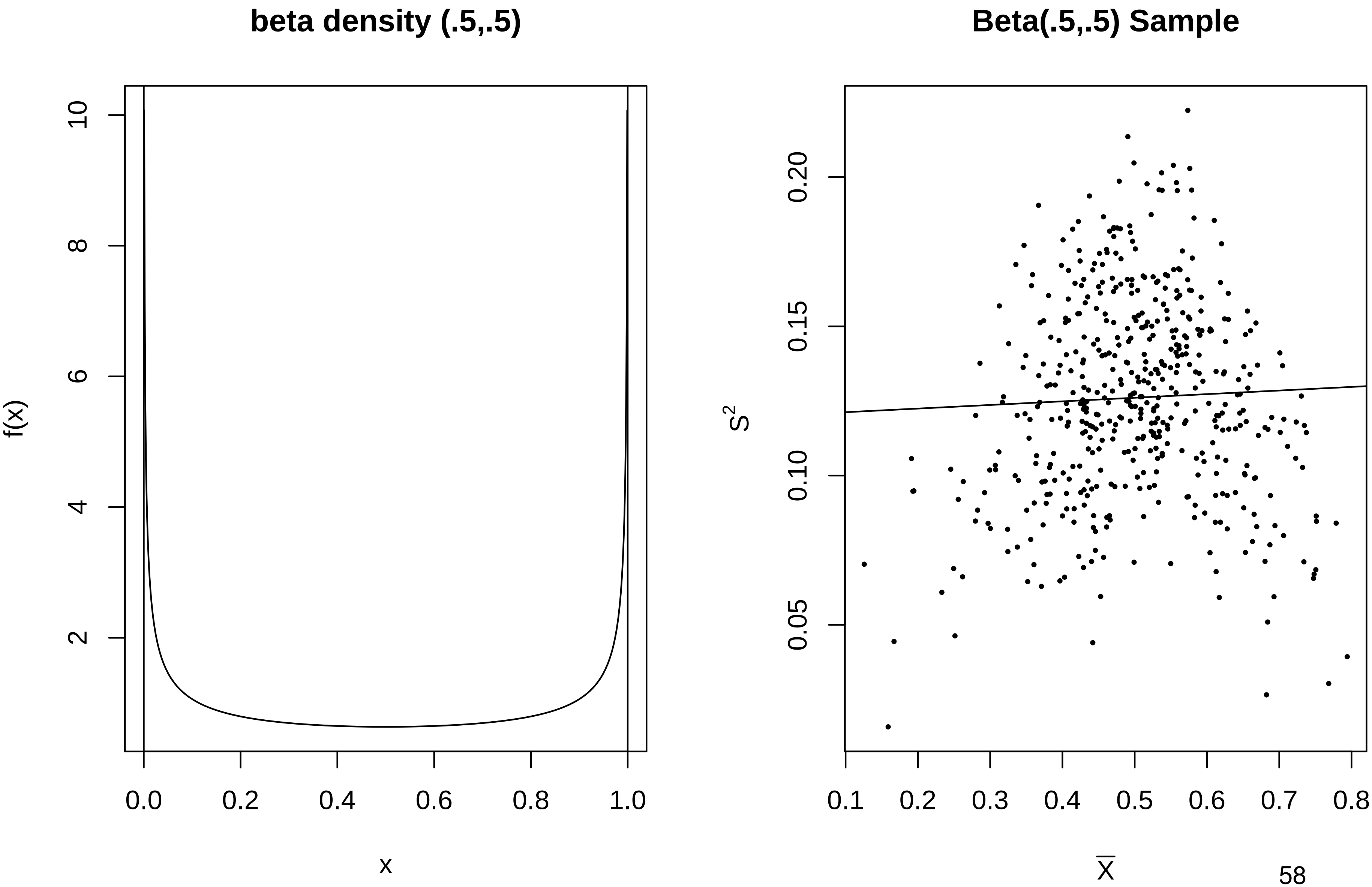
It appears that for symmetric sampled distribution (normal, uniform, and t) the least squares line, fitting S^2 as linear function of $\bar{X}$, is roughly horizontal, indicating zero correlation, but not necessarily independence.

For uniform samples a high values of S^2 seem to be associated with an $\bar{X}$ near .5.

For t-samples large S^2 values seem to indicate outliers which also affect $\bar{X}$.

When the sampled distribution is skewed (Chi-Square, Weibull, Poisson) the least squares line shows a definite slope, i.e., we have correlation and definitely dependence.

$\bar{X}$ and S^2 Dependence (500 Samples of Size $n = 10$)



Comments

The sampled beta density over the interval $(0,1)$ is shown on the left.

It gives higher density near 0 or 1 than in the middle.

The $(\bar{X}, S^2)$ scatter shows a definite wedge pattern of high values of S^2 being associated with more central values of $\bar{X}$.

Values of $\bar{X}$ near 0 or 1 are associated with low values of S^2.

These associations make sense. For example, a value of $\bar{X}$ near 1 can only come about when most observations are on the right side of .5, leading to a smaller S^2.

We clearly see dependence in spite of zero correlation patterns.

One-Sample t-Test

Assume that $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

We want to test the hypothesis $H_0 : \mu = \mu_0$ against the alternatives $H_1 : \mu \neq \mu_0$.

σ is left unspecified and is unknown. H_0 is a composite hypothesis.

$\bar{X}$ is a good indicator for μ since its mean is μ and its variance is $\sigma^2(\bar{X}) = \sigma^2/n$.

Thus a reasonable test statistic may be $\bar{X} - \mu_0 \sim \mathcal{N}(\mu - \mu_0, \sigma^2/n) = \mathcal{N}(0, \sigma^2/n)$

The last $=$ holds when H_0 is true. Unfortunately we do not know σ.

$\sqrt{n}(\bar{X} - \mu_0)/\sigma = (\bar{X} - \mu_0)/(\sigma/\sqrt{n}) \sim \mathcal{N}(0,1)$ suggests replacing the unknown σ

by suitable estimate to get a single reference distribution under H_0.

From the previous slide: $\Longrightarrow$ $s^2 = \sum_{i=1}^n (X_i - \bar{X})^2/(n-1) \sim \sigma^2 C_{n-1}/(n-1)$

s^2 is independent of $\bar{X}$. Note $E(s^2) = \sigma^2$, i.e., s^2 is an unbiased estimate of σ^2.

One-Sample t-Statistic

Replacing σ by s in the standardization $\sqrt{n}(\bar{X}-\mu_0)/\sigma \Longrightarrow$ one-sample t-statistic

$$t(\mathbf{X}) = \frac{(\bar{X}-\mu_0)}{s/\sqrt{n}} = \frac{\sqrt{n}(\bar{X}-\mu_0)/\sigma}{\sqrt{s^2/\sigma^2}} = \frac{\sqrt{n}(\bar{X}-\mu_0)/\sigma}{\sqrt{C_{n-1}/(n-1)}} = \frac{Z}{\sqrt{C_{n-1}/(n-1)}} \sim t_{n-1}$$

since under H_0 we have that $Z = \sqrt{n}(\bar{X}-\mu_0)/\sigma \sim \mathcal{N}(0,1)$ and $C_{n-1} \sim \chi^2_{n-1}$, both independent of each other. We thus satisfy the definition of the t-distribution.

Hence we can use $t(\mathbf{X})$ in conjunction with the single known reference distribution t_{n-1} under the composite hypothesis H_0 and reject H_0 for large values of $|t(\mathbf{X})|$.

The 2-sided level α test has critical value

$t_{\text{crit}} = t_{n-1,1-\alpha/2} = \texttt{qt}(1-\alpha/2,\texttt{n}-1) = \texttt{t.crit}$.

We reject H_0 when $|t(\mathbf{X})| \geq t_{\text{crit}}$.

The 2-sided p-value for the observed t-statistic $t_{\text{obs}}(\mathbf{x}) = \texttt{t.obs}$ is

$P(|t_{n-1}| \geq |t_{\text{obs}}(\mathbf{x})|) = 2P(t_{n-1} \leq -|t_{\text{obs}}(\mathbf{x})|) = 2 * \texttt{pt}(-\texttt{abs}(\texttt{t.obs}),\texttt{n}-1)$.

The t.test in R

R has a function, `t.test`, that performs 1- and 2-sample t-tests.

See `?t.test` for documentation. We focus here on the 1-sample test.

```
> t.test(rnorm(20)+.4)

        One Sample t-test

data:  rnorm(20) + 0.4
t = 2.2076, df = 19, p-value = 0.03976
alternative hypothesis: true mean is not equal to 0
95 percent confidence interval:
 0.02248992 0.84390488
sample estimates:
mean of x
0.4331974
```

Calculation of the Power Function of the Two-Sided t-Test

The power function of this two-sided t-test is given by

$$\beta(\mu,\sigma) = P\left(|t| \geq t_{\text{crit}}\right) = P(t \leq -t_{\text{crit}}) + P(t \geq t_{\text{crit}}) = P(t \leq -t_{\text{crit}}) + 1 - P(t < t_{\text{crit}})$$

$$\begin{aligned} t = t(\mathbf{X}) &= \frac{\sqrt{n}(\bar{X}-\mu_0)}{s} = \frac{\sqrt{n}(\bar{X}-\mu+(\mu-\mu_0))/\sigma}{s/\sigma} \\ &= \frac{\sqrt{n}(\bar{X}-\mu)/\sigma + \sqrt{n}(\mu-\mu_0)/\sigma}{s/\sigma} = \frac{Z+\delta}{\sqrt{C_{n-1}/(n-1)}} \sim t_{n-1,\delta} \end{aligned}$$

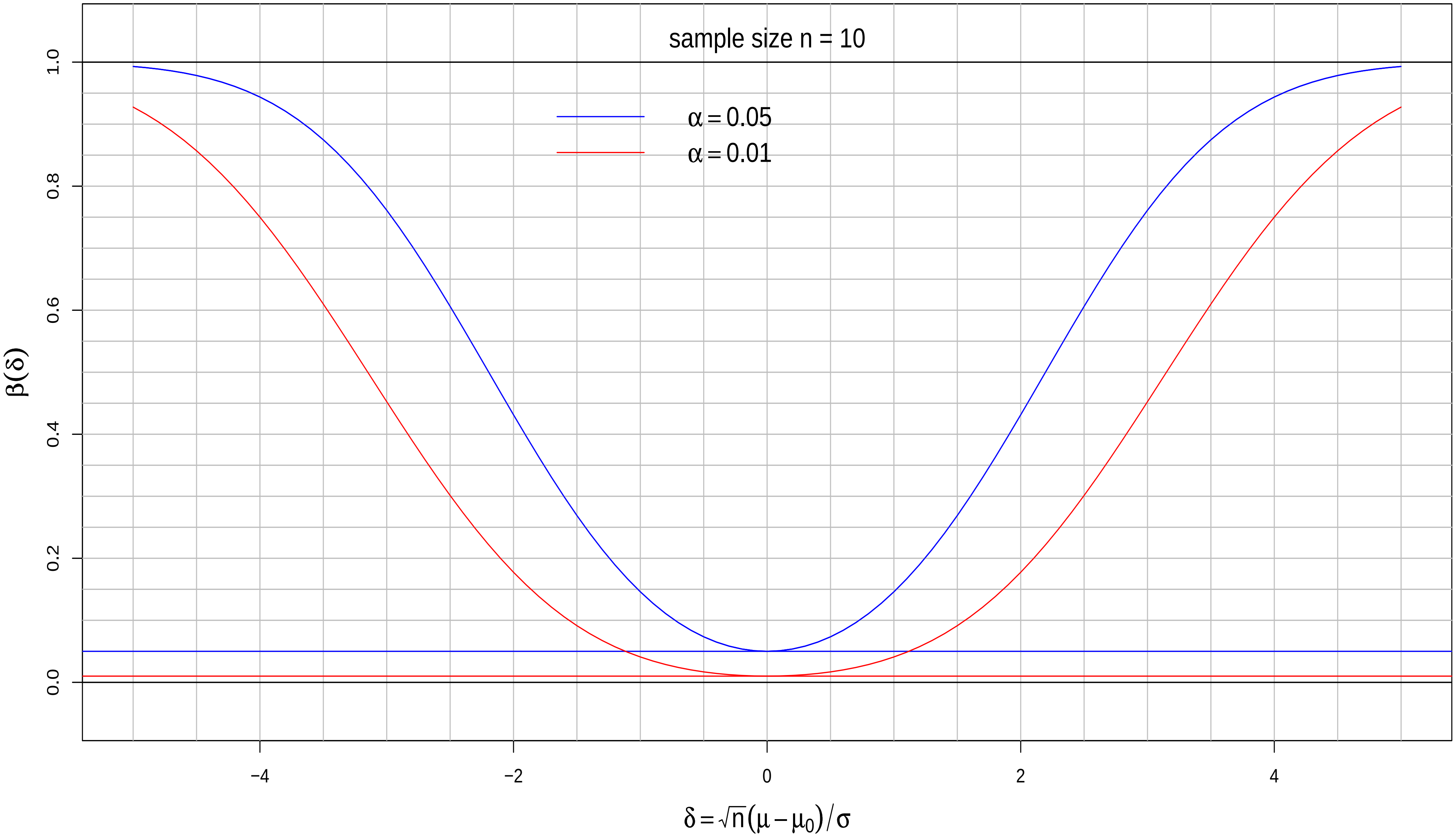
noncentral t-distribution with noncentrality parameter $\delta = \sqrt{n}(\mu-\mu_0)/\sigma = \texttt{delta}$.

Thus the power function depends on μ and σ only through δ and we write

$$\begin{aligned} \beta(\delta) &= P(t_{n-1,\delta} \leq -t_{\text{crit}}) + 1 - P(t_{n-1,\delta} < t_{\text{crit}}) \\ &= \texttt{pt}(-\texttt{t.crit},\texttt{n}-1,\texttt{delta}) + 1 - \texttt{pt}(\texttt{t.crit},\texttt{n}-1,\texttt{delta}) \end{aligned}$$

The power function also depends on n.

Power Function of Two-Sided t-Test



How to Use the Power Function

For the level $\alpha = .05$ test in the previous plot we can read off

$$\beta(\delta) \approx .6 \quad \text{for} \quad \delta = \pm\sqrt{n}(\mu_0 - \mu)/\sigma \approx 2.5 \quad \text{or} \quad |\mu_0 - \mu| \approx 2.5\sigma/\sqrt{n}.$$

The smaller the natural variability σ the smaller the difference $|\mu_0 - \mu|$ that we can detect with probability $.6$.

Similarly, the larger the sample size n the smaller the difference $|\mu_0 - \mu|$ we can detect with probability $.6$, note however the square root effect in $\sqrt{n}$.

Both of these conclusions are intuitive because $\sigma(\bar{X}) = \sigma/\sqrt{n}$.

Given a required detection difference $|\mu - \mu_0|$ and with some upper bound knowledge $\sigma_u \geq \sigma$ we can plan the appropriate minimum sample size n to achieve the desired power $\beta = .6$: $\quad 2.5 \times \sigma/|\mu - \mu_0| \leq 2.5 \times \sigma_u/|\mu - \mu_0| = \sqrt{n}$.
For power $\neq .6$ replace 2.5 by the appropriate value from the previous plot.

Where is the Flaw in Previous Argument?

We tacitly assumed that the power curve plot would not change with n, i.e., we consider the effect of n only via $\delta = \sqrt{n}\,|\mu - \mu_0|/\sigma$ on the plot abscissa.

Both $t_{\text{crit}} = \texttt{qt}(1-\alpha/2, \texttt{n}-1)$ and $P(t_{n-1,\delta} \le \pm t_{\text{crit}})$ depend on n, as does $\delta = \sqrt{n}\,|\mu - \mu_0|/\sigma$. See the next 3 plots.

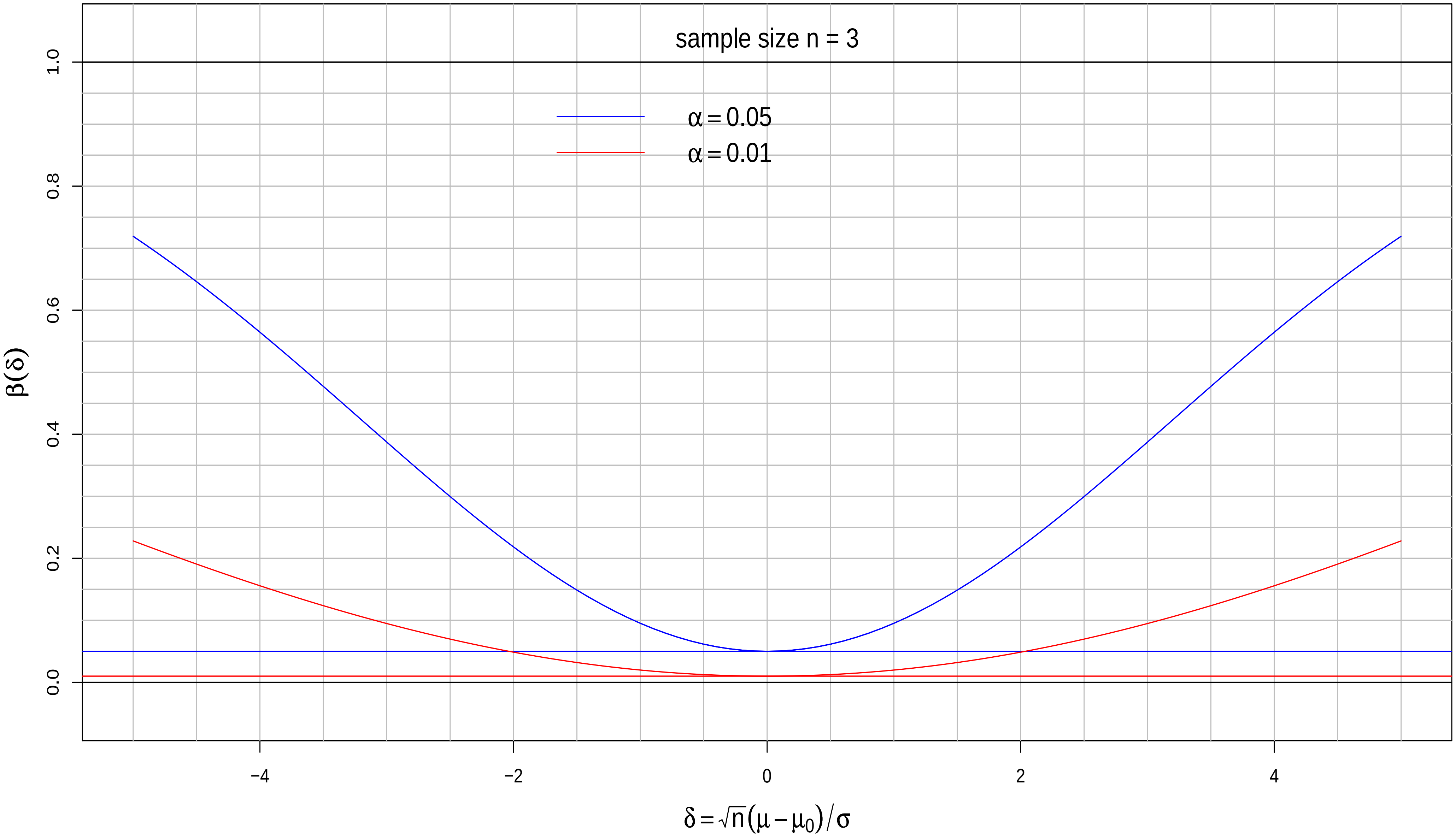
Thus it does not suffice to consider the n in δ alone.

However, typically the sample size requirements will ask for large values of n.

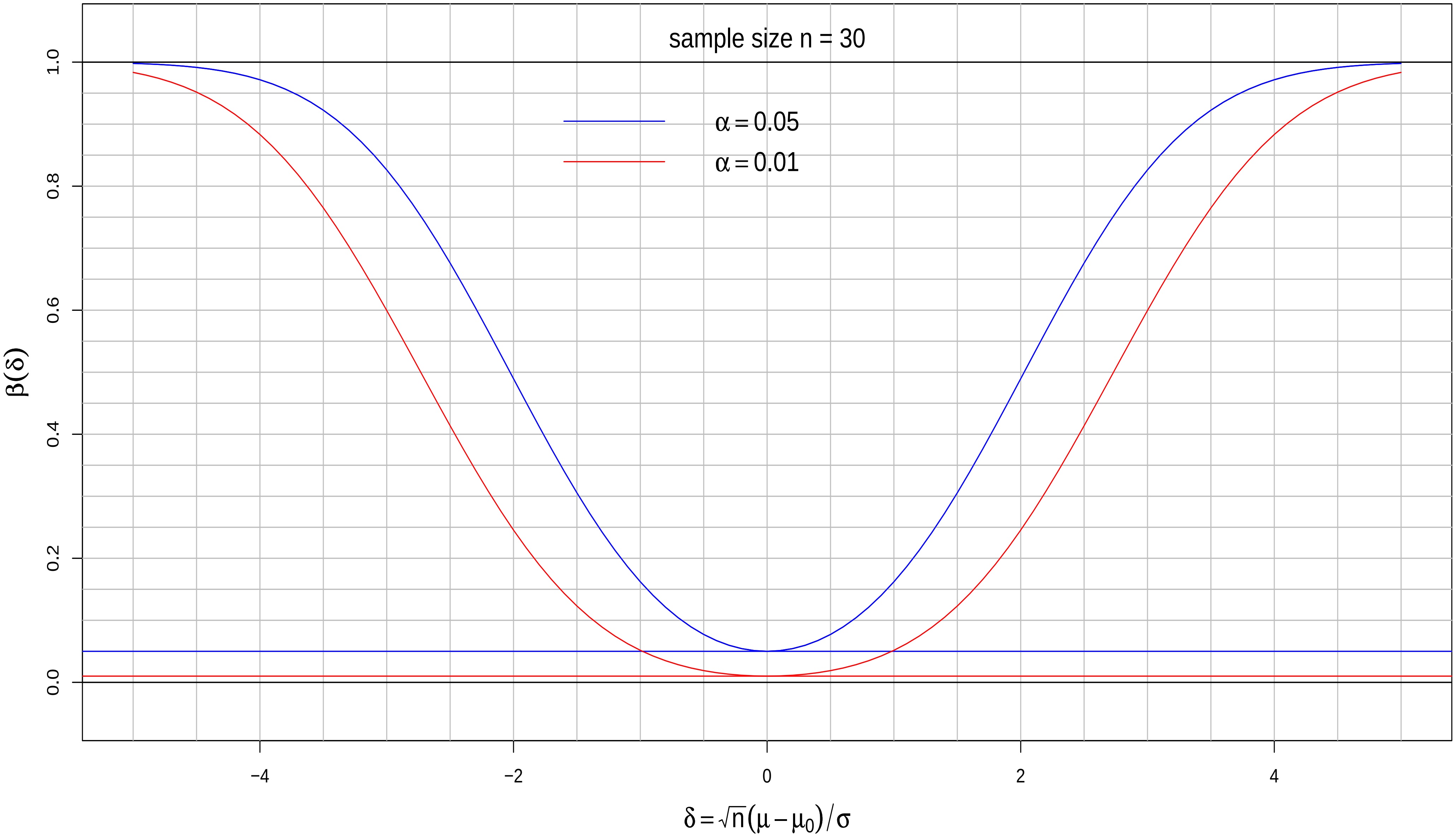
In that case $t_{\text{crit}} \approx \texttt{qnorm}(1-\alpha/2)$ and $t_{n-1,\delta} \approx \mathcal{N}(\delta, 1)$ stabilize (for fixed δ). Compare $n = 100$ and $n = 1000$ in the next few plots. For large n, most of the benefit from increasing n comes via increasing $\delta = \sqrt{n}\,|\mu - \mu_0|/\sigma$.

We will provide a function that gets us out of this dilemma.

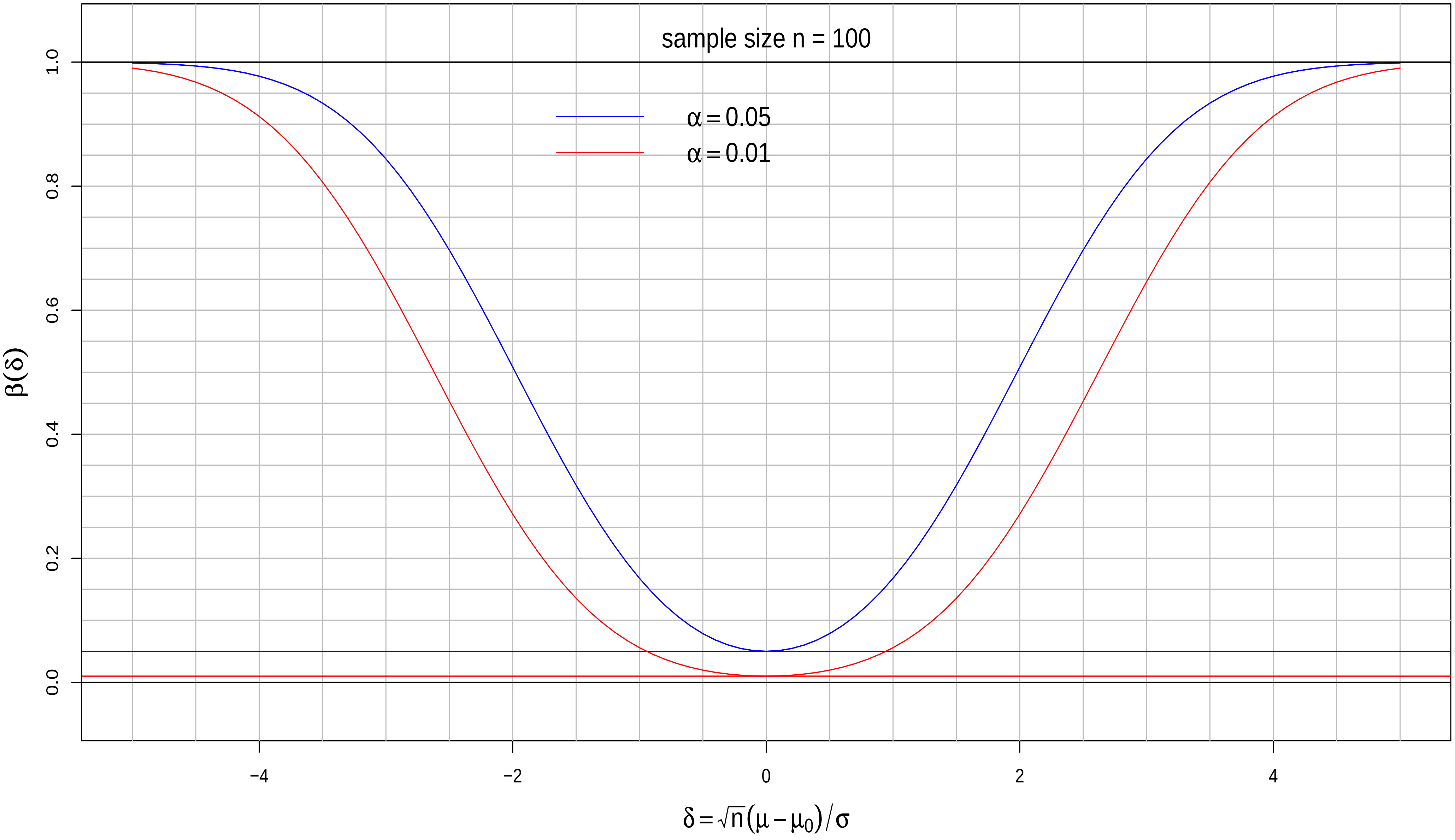
Power Function of Two-Sided t-Test



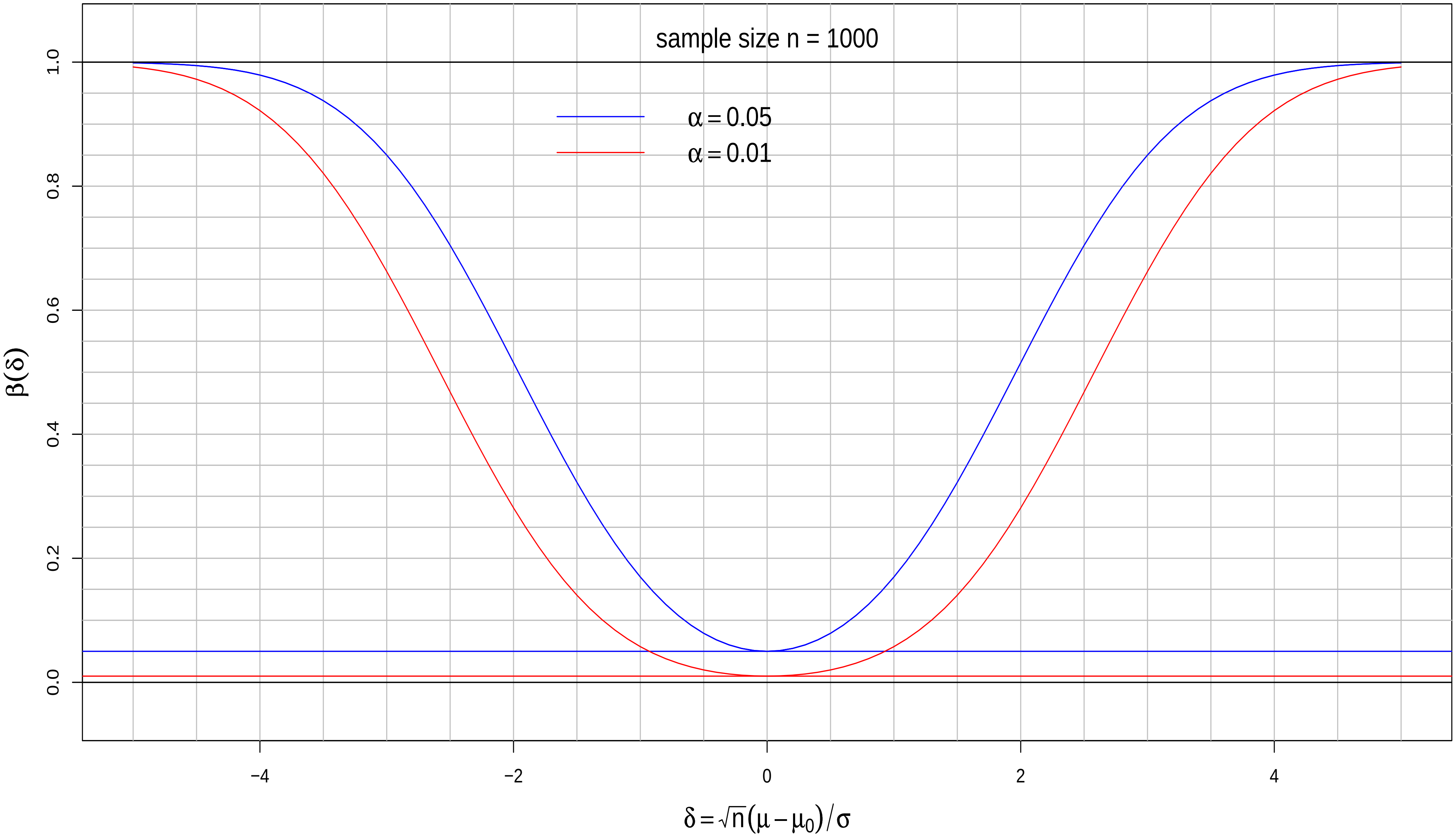
Power Function of Two-Sided *t*-Test



Power Function of Two-Sided t-Test



Power Function of Two-Sided t-Test



Some Discussion of the Impact of n

The previous slides showed that the impact of n on the power function curve shapes can be substantial for small n (see $n = 3 \quad \Longrightarrow \quad n = 30$).

However, for larger n the power functions only change their shape slightly, see the small change as $n = 30 \rightarrow n = 100 \rightarrow n = 1000$.

In those (larger n) cases the main impact of n on power is through the noncentrality parameter $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$.

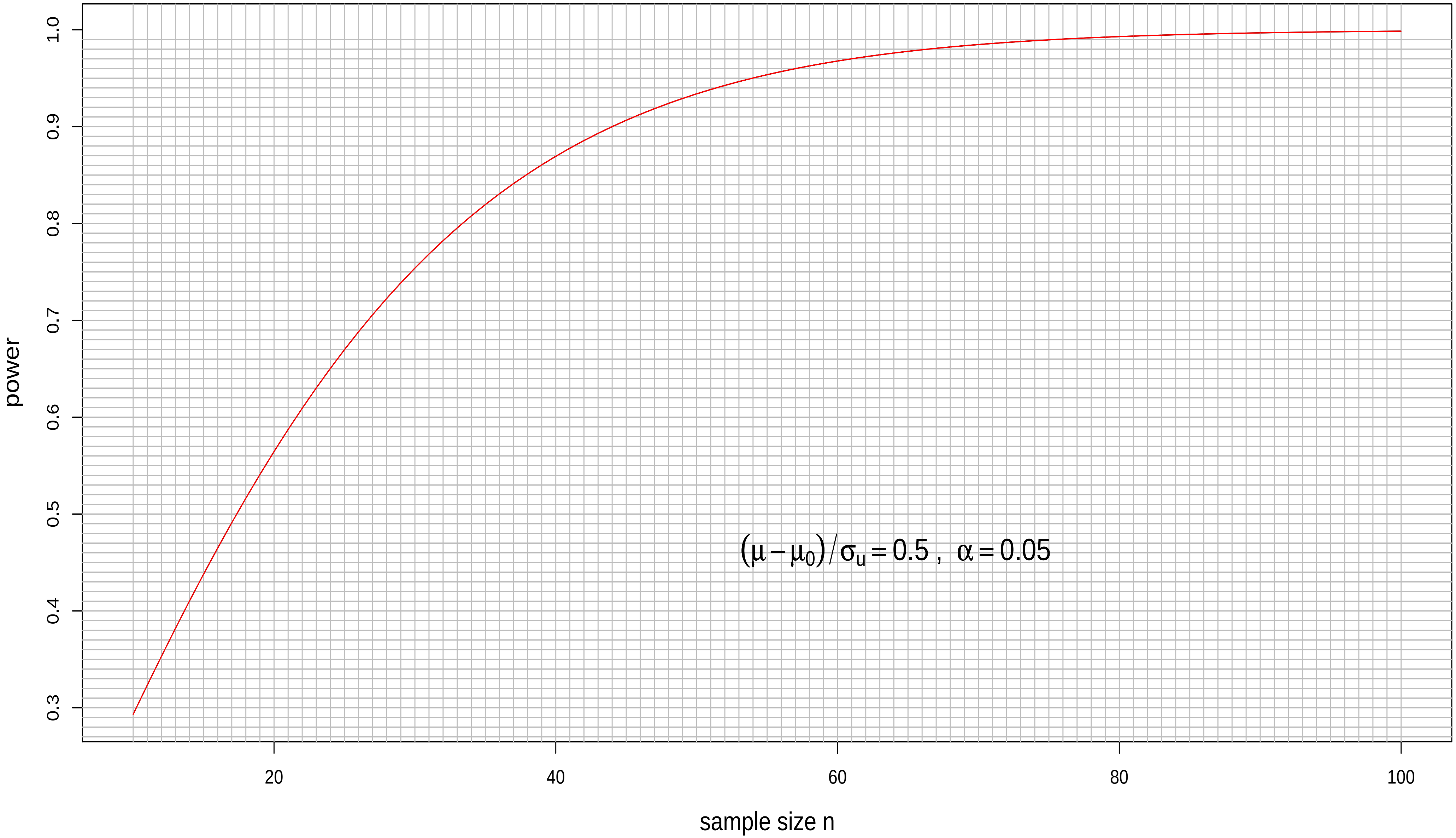
Sample Size Function (2-sided)

```
sample.size2 = function(delta0=1,nrange=10:100,alpha=.05){
power=NULL
for(n in nrange){
  tcrit=qt(1-alpha/2,n-1)
  power=c(power,1-pt(tcrit,n-1,sqrt(n)*delta0)+
                           pt(-tcrit,n-1,sqrt(n)*delta0))
}
plot(nrange,power,type="l",xlab="sample size n")
abline(h=seq(.01,.99,.01),col="grey")
abline(v=nrange,col="grey")
title(substitute((mu-mu[0])/sigma[u]==delta0,list(delta0=delta0))) }
```

Here `delta0` $= (\mu - \mu_0)/\sigma_u$ is the point at which we want to achieve a given power. σ_u is a known (conservative) upper bound on σ.

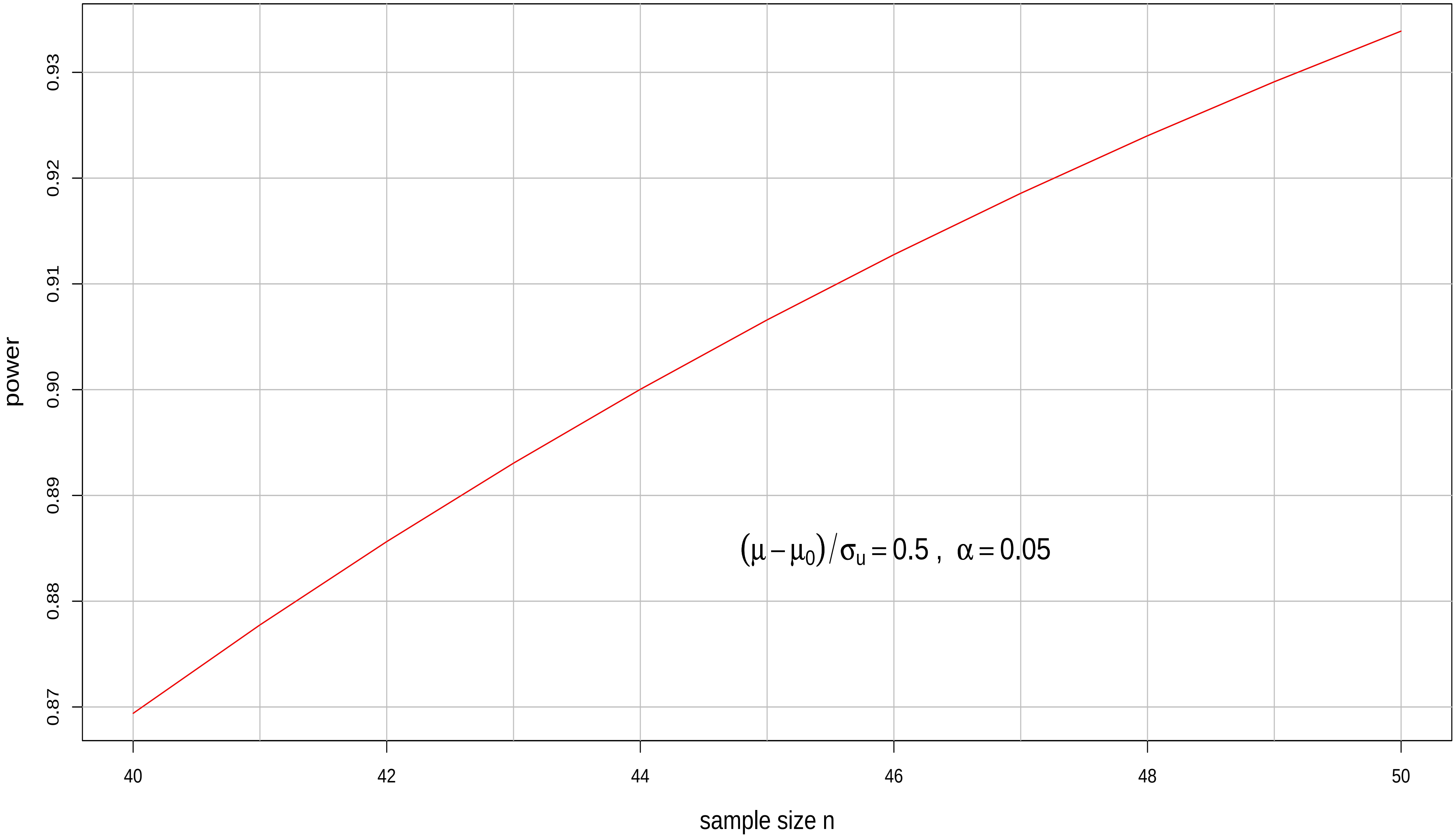
`nrange` gives a vector of sample sizes at which the power function is computed for `delta0` and significance level `alpha`.

Power of Two-Sided t-Test for Various n



Want power $.9$ at $\mu - \mu_0 = .5\sigma_u$.

Power of Two-Sided t-Test for Various n (refined view)



$n = 44$ will do for power .9 at $\mu - \mu_0 = .5\sigma_u$.

The Blessings of $(\mu-\mu_0)/\sigma$

The power function depends on $\mu-\mu_0$ only in relation to σ units.

This is a sensible benefit.

If σ is large, only very large differences between μ and μ_0 would matter.

Smaller differences $\mu-\mu_0$ would get swamped or appear irrelevant when compared with the population variation, i.e., σ.

One-Sided One-Sample Testing Problem

Again assume $\mathbf{X} = (X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

In some situations we may want to test $H_0' : \mu = \mu_0$ against $H_1 : \mu > \mu_0$.

More broadly, we may want to test $H_0 : \mu \leq \mu_0$ against $H_1 : \mu > \mu_0$.

Clearly, large values of $\bar{X} - \mu_0$ speak for H_1 and against H_0' or H_0.

Large negative values of $\bar{X} - \mu_0$ may also speak against H_0' but not against H_0.

Thus we should be very clear as to the hypotheses being tested,

and whether it should be a one-sided or two-sided alternative.

One-Sided One-Sample t-Test

Based on the discussion for the 2-sided testing problem it is natural to consider

$$t = t(\mathbf{X}) = \frac{\bar{X} - \mu_0}{s/\sqrt{n}}$$

as our test statistic and reject H_0 in favor of H_1 when t is too large.

What is the reference distribution of t under $H_0 : \mu \le \mu_0$?

Recall that $t \sim t_{n-1,\delta}$, the noncentral t-distribution with noncentrality parameter $\delta = \sqrt{n}(\mu - \mu_0)/\sigma$.

Note that $\mu \le \mu_0 \quad \Longleftrightarrow \quad \delta = \sqrt{n}(\mu - \mu_0)/\sigma \le 0.$

Thus we have many different reference distributions under $H_0 : \delta \le 0.$

Power Function of the One-Sided One-Sample t-Test

Suppose we reject H_0 when $t \geq t_{\text{crit}}$ for some chosen critical value t_{crit}.

Then the power function again depends only on δ and n and is given by

$$\beta(\delta) = P(t_{n-1,\delta} \geq t_{\text{crit}}) = 1 - \texttt{pt}(\texttt{t}_{\text{crit}}, \texttt{n} - 1, \delta).$$

$\beta(\delta)$ is strictly increasing in δ since

$$P\left(\frac{Z+\delta}{\sqrt{C_{n-1}/(n-1)}} \geq t_{\text{crit}}\right) = P\left(Z \geq t_{\text{crit}}\sqrt{C_{n-1}/(n-1)} - \delta\right) \nearrow \quad \text{as } \delta \nearrow .$$

Thus $\beta(\delta) \leq \beta(0)$ for $\delta \leq 0$ and we have $\max_{\delta \leq 0}\{\,\beta(\delta)\,\} = \beta(0)$.

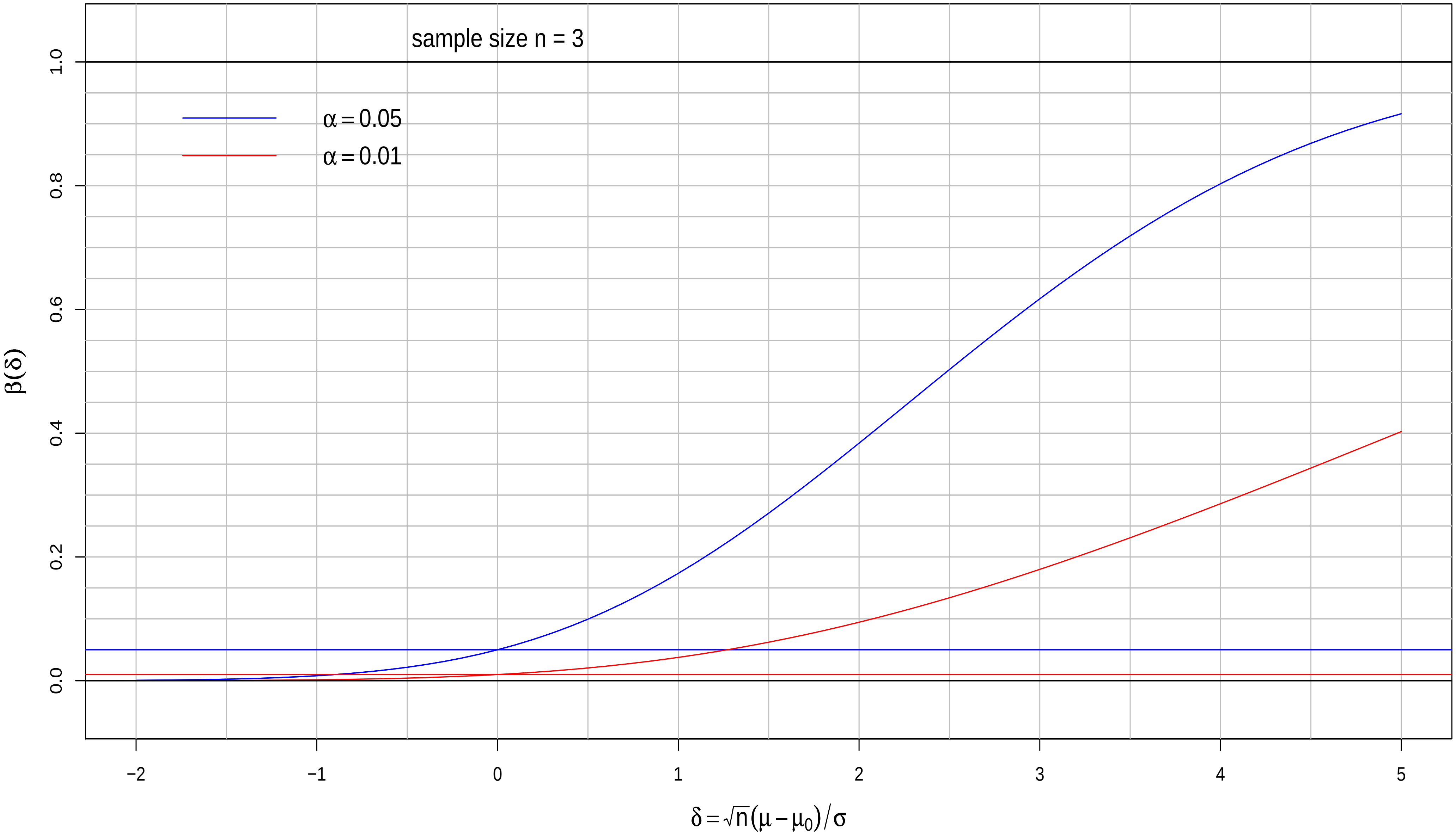
Thus choose $t_{\text{crit}} = \texttt{t.crit}$ such that $\beta(0) = \alpha$, i.e,

$$\alpha = \beta(0) = P(t_{n-1,0} \geq t_{\text{crit}}) = P(t_{n-1} \geq t_{\text{crit}}) \quad \text{or} \quad \texttt{t.crit} = \texttt{qt}(1 - \alpha, \texttt{n} - 1)\ .$$

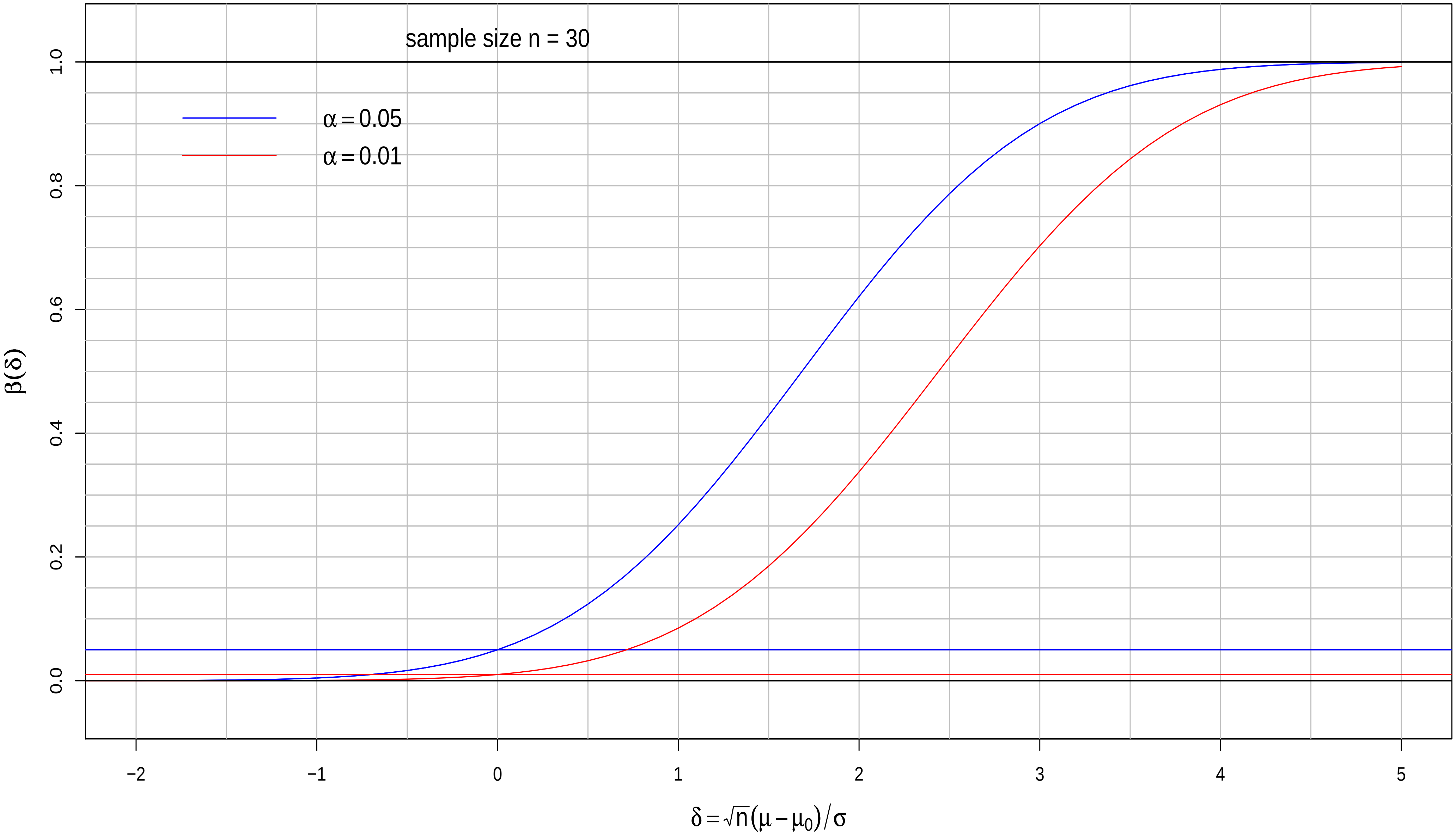
Using t.test

```
> t.test(rnorm(20)+.4,alternative="greater")

        One Sample t-test

data:  rnorm(20) + 0.4
t = 2.4646, df = 19, p-value = 0.01171
alternative hypothesis: true mean is greater than 0
95 percent confidence interval:
 0.1429615       Inf
sample estimates: mean of x 0.4790822
```

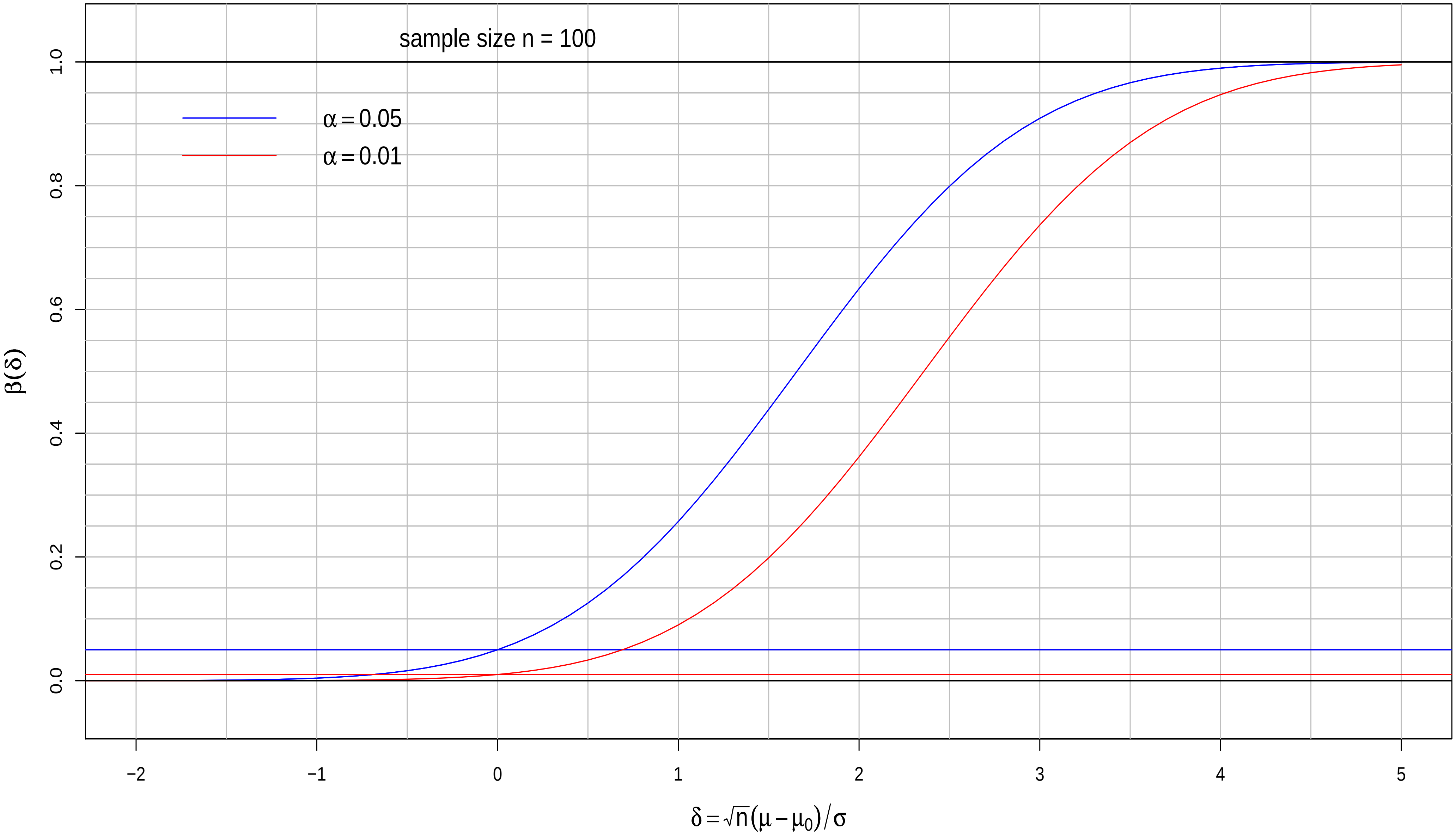
Power Function of One-Sided *t*-Test



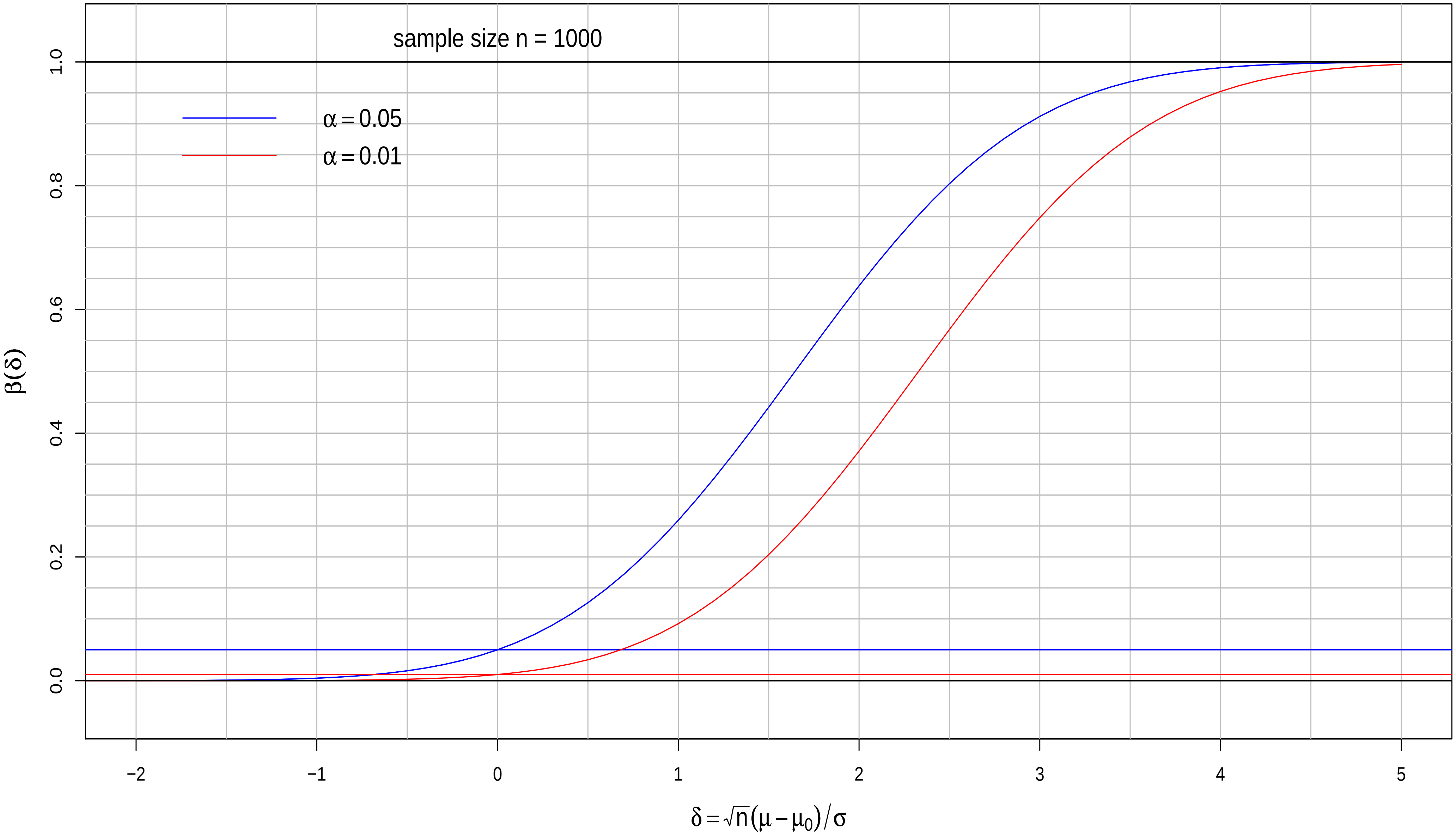
Power Function of One-Sided *t*-Test



Power Function of One-Sided *t*-Test



Power Function of One-Sided *t*-Test



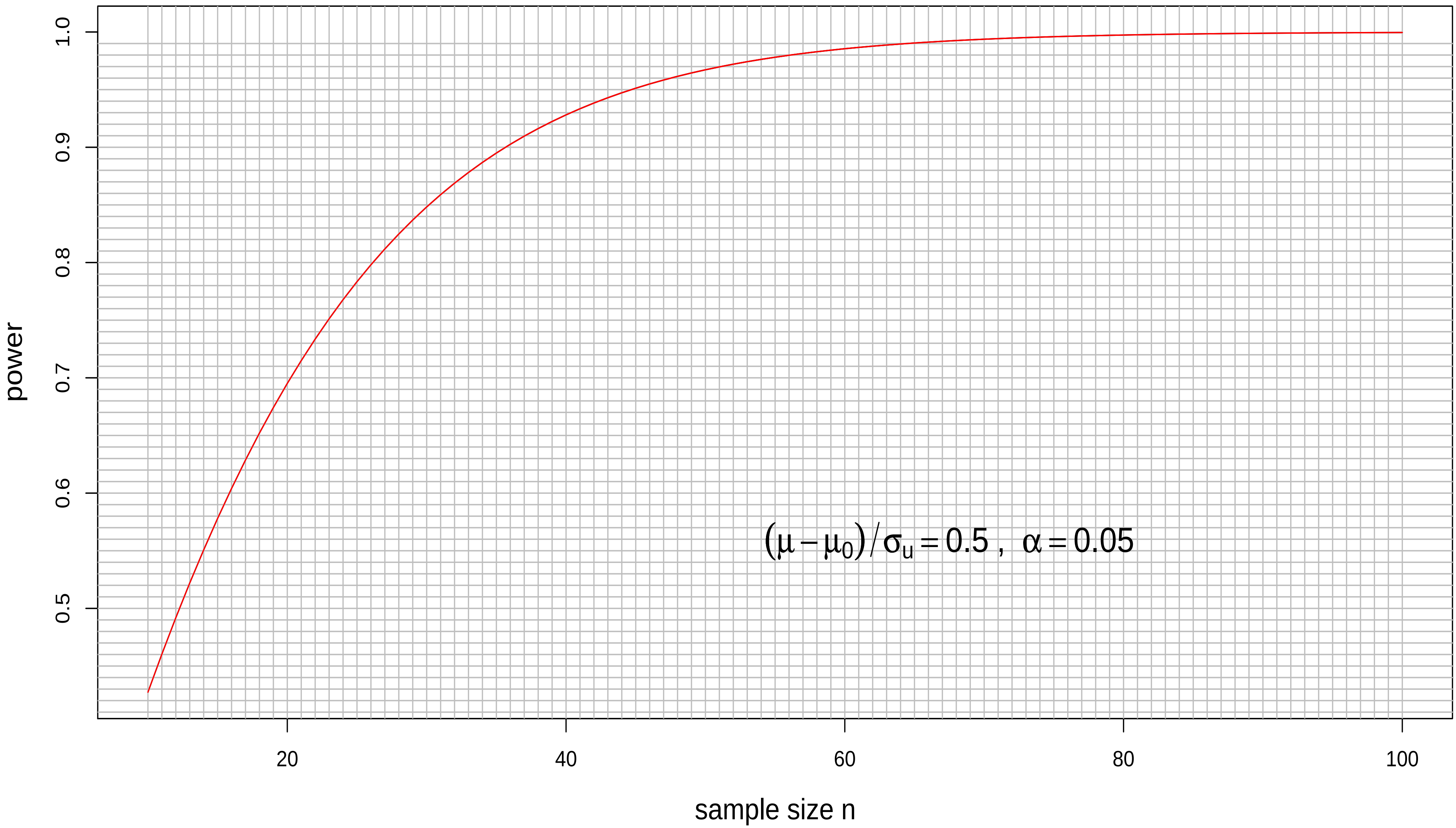
Sample Size Function (1-sided)

```
sample.size1 = function(delta0=1,nrange=10:100,alpha=.05){
power=NULL
for(n in nrange){
 tcrit=qt(1-alpha,n-1)
 power=c(power,1-pt(tcrit,n-1,sqrt(n)*delta0)) }
plot(nrange,power,type="l",xlab="sample size n")
abline(h=seq(.01,.99,.01),col="grey")
abline(v=nrange,col="grey")
title(substitute((mu-mu[0])/sigma[u]==delta0~
","~alpha==alpha0,list(delta0=delta0,alpha0=alpha)))
lines(nrange,power,col="red") }
```

Here `delta0` $=(\mu-\mu_0)/\sigma_u$ is the point at which we want to achieve a given power. σ_u is a known (or assumed conservative) upper bound on σ.

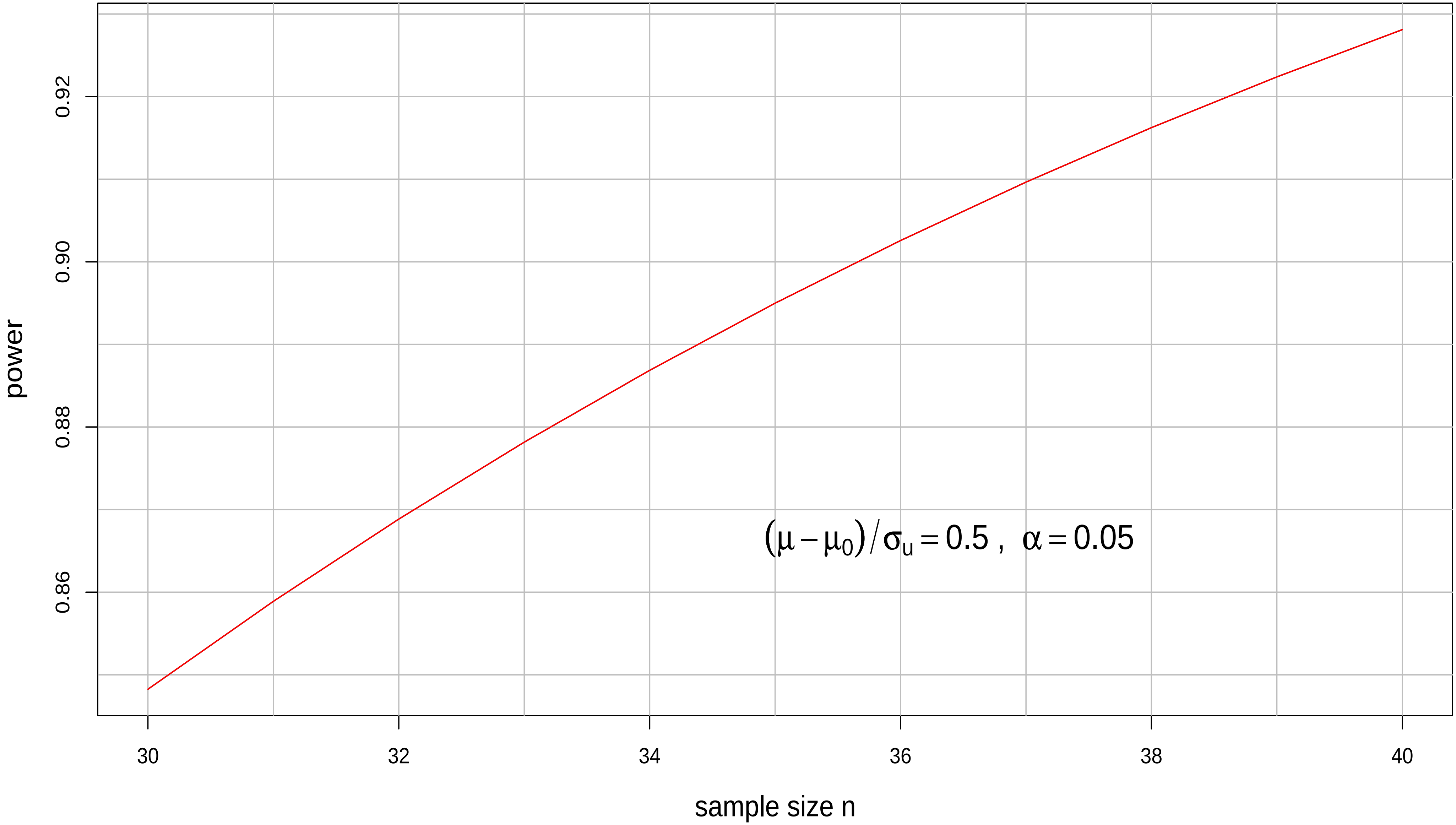
`nrange` gives a vector of sample sizes at which the power function is computed for `delta0` and significance level `alpha`.

Power of One-Sided t-Test for Various n



Want power $.9$ at $\mu - \mu_0 = .5\sigma_u$.

Power of One-Sided t-Test for Various n (refined view)



$n = 36$ will do for power $.9$ at $\mu - \mu_0 = .5\sigma_u$.

Hypothesis Tests & Confidence Intervals

For testing $H_0 : \mu = \mu_0$ we accept H_0 with the two-sided t-test whenever

$$\left|\frac{\bar{X} - \mu_0}{s/\sqrt{n}}\right| < t_{n-1,1-\alpha/2} \quad \Longleftrightarrow \quad \mu_0 \in \bar{X} \pm t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}$$

Thus the interval $\bar{X} \pm t_{n-1,1-\alpha/2} \times s/\sqrt{n}$ consists of all acceptable μ_0,

i.e., all μ_0 for which one would accept $H_0 : \mu = \mu_0$ at level α.

Furthermore, since under H_0 our acceptance probability is $1 - \alpha$ we have

$$P_{\mu_0}\left(\bar{X} - t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}} \;<\; \mu_0 \;<\; \bar{X} + t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

Here the subscript μ_0 on P indicates the assumed true value of the mean μ. Since this holds for any value μ_0 we may as well drop the subscript 0 on μ_0 and write

$$P_{\mu}\left(\bar{X} - t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}} \;<\; \mu \;<\; \bar{X} + t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}\right) = 1 - \alpha$$

The choice of $<$ or $\leq$ is immaterial. The case “$=$” has probability zero!

The Nature of Confidence Intervals

$$\left[\bar{X} - t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}\ ,\ \ \bar{X} + t_{n-1,1-\alpha/2} \times \frac{s}{\sqrt{n}}\right]$$

is called a $100 \times (1-\alpha)\%$ confidence interval for the unknown mean μ.

It is a random interval which has probability $1-\alpha$ of covering μ.

This is not a statement about μ being random due to being unknown or uncertain.

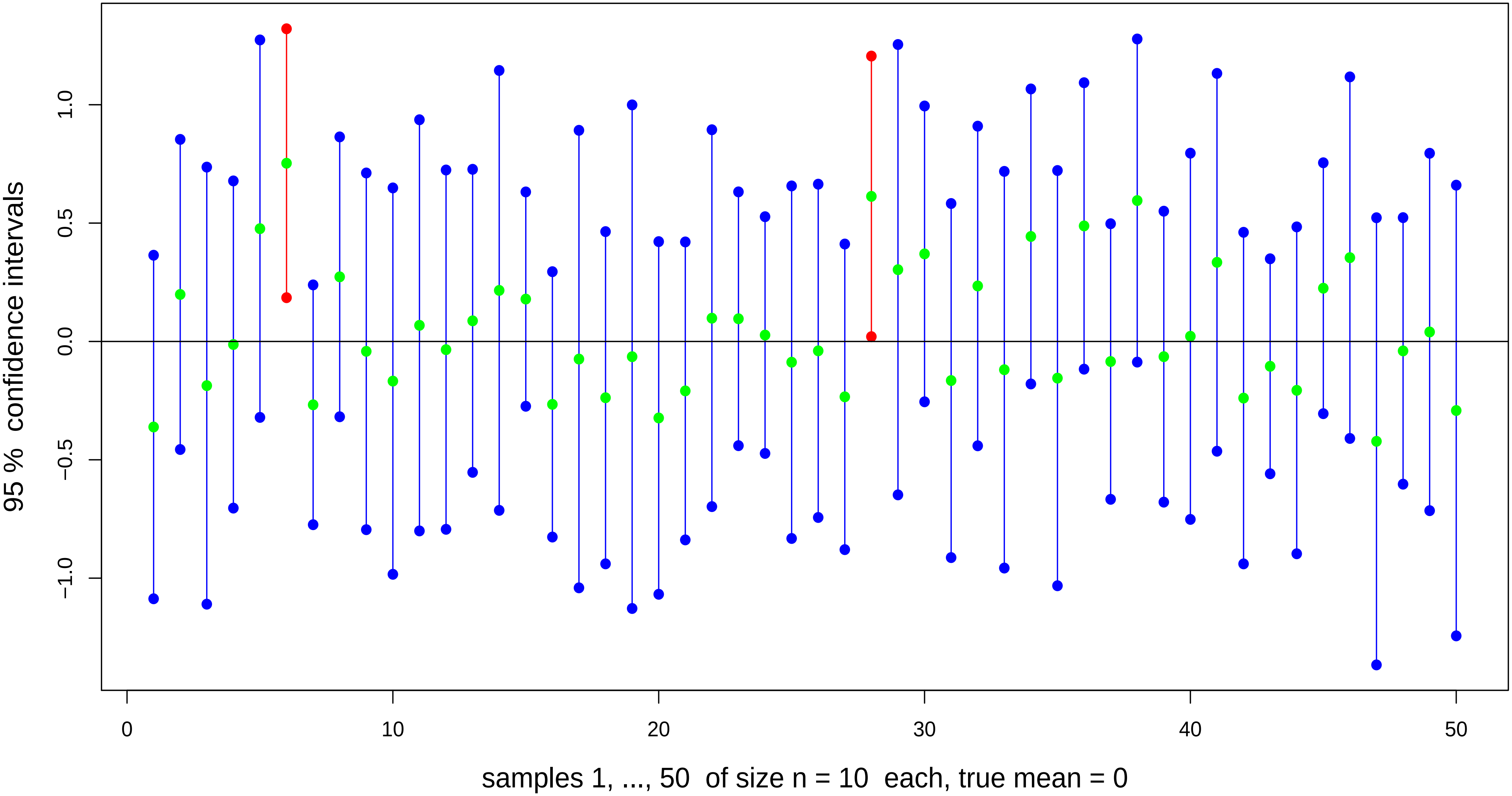
μ does not "fall" into that interval with probability $1-\alpha$. μ is fixed but unknown.

Without knowing μ we will not know whether the interval covers μ or not.

"Statistics means never having to say you're certain." (Myles Hollander)

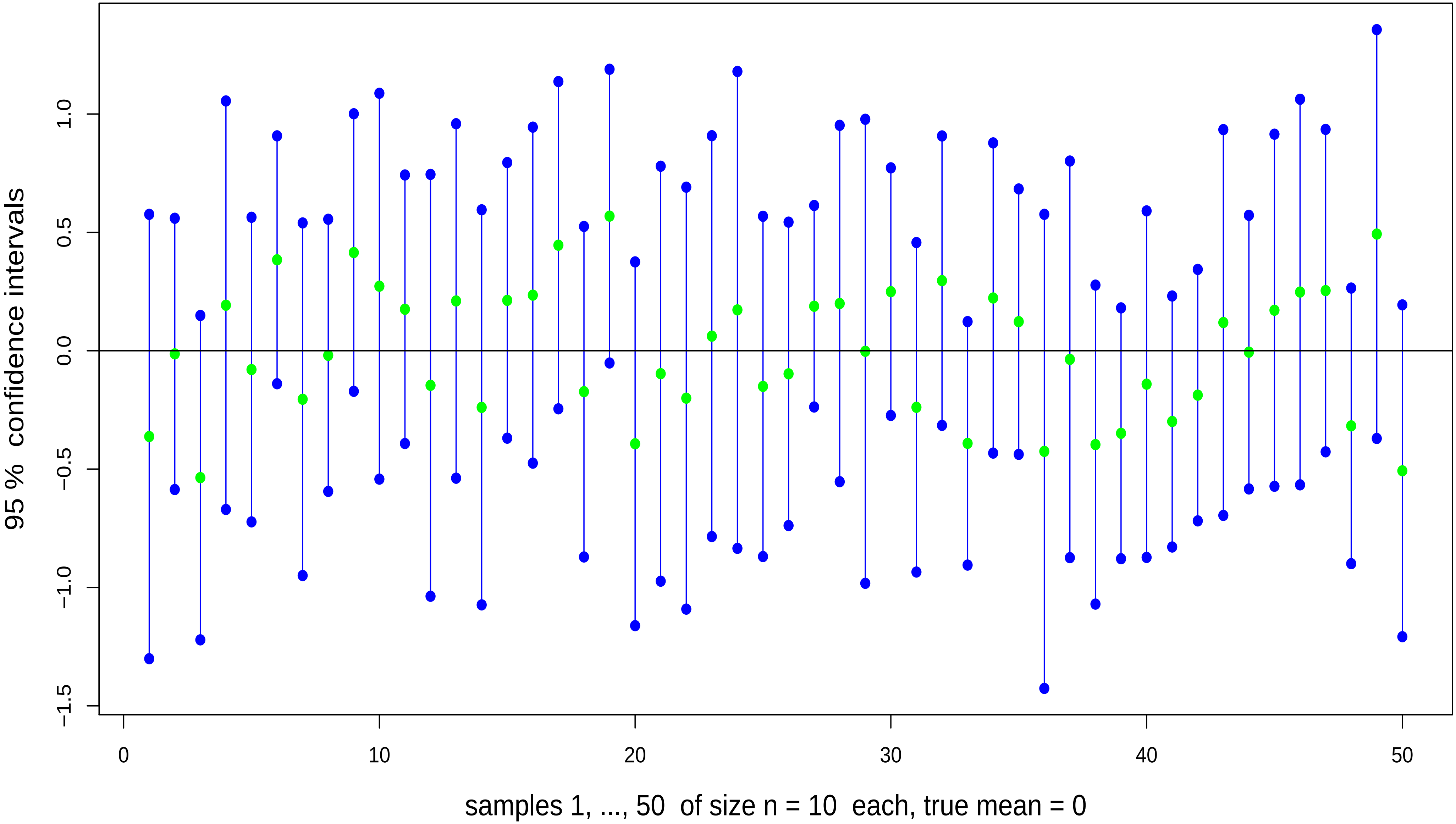
$\Longleftarrow$ "Love means never having to say you're sorry," Love Story by Eric Segal

50 Confidence Intervals

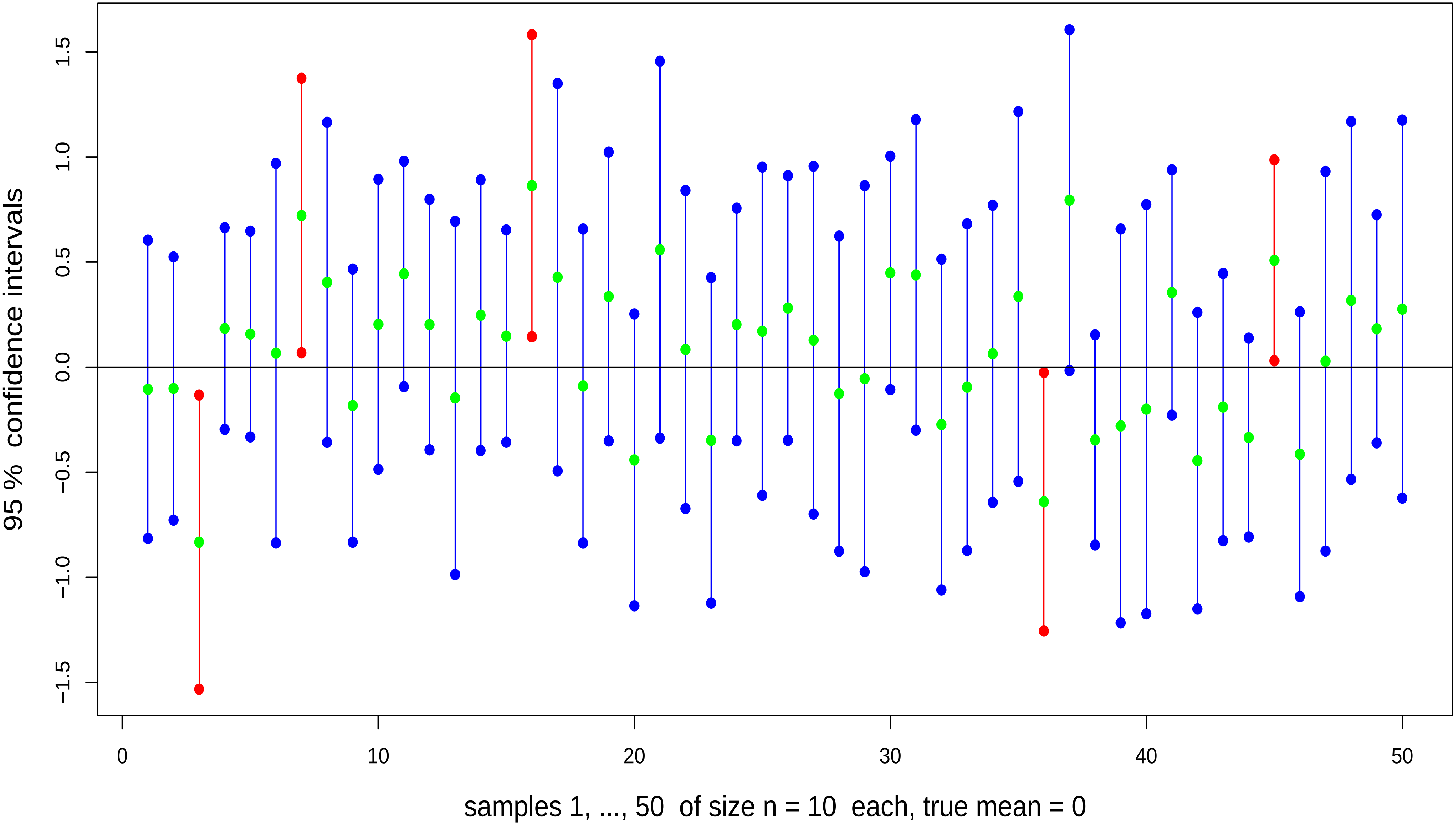


Expected Missed Coverages $50 \times .05 = 2.5$

50 Confidence Intervals



50 Confidence Intervals



Using Confidence Intervals to Test Hypotheses

Not only do confidence intervals provide a more informative way of estimating parameters, as opposed to just stating the interval midpoint $\bar{X}$ as estimate for μ, they can also be used to directly test hypotheses.

No surprise: Confidence intervals were introduced via acceptable hypotheses.

Thus we reject $H : \mu = \mu_0$ at level α whenever the $100(1-\alpha)\%$ confidence interval does not cover μ_0.

If the coverage probability is at least $1-\alpha$, then the probability of missing the true target (rejecting falsely) is at most α

What about One-Sided Hypotheses?

Testing $H : \mu \le \mu_0$ against $H_1 : \mu > \mu_0$ we reject when $\sqrt{n}(\bar{X} - \mu_0)/s \ge t_{n-1,1-\alpha}$, or accept whenever

$$\sqrt{n}(\bar{X} - \mu_0)/s < t_{n-1,1-\alpha} \qquad \Longleftrightarrow \qquad \mu_0 > \bar{X} - t_{n-1,1-\alpha} \times s/\sqrt{n} \, ,$$

i.e., $\bar{X} - t_{n-1,1-\alpha} \times s/\sqrt{n}$ is a $100(1-\alpha)\%$ lower confidence bound for μ_0.

We could also state it in open ended interval form: $(\bar{X} - t_{n-1,1-\alpha} \times s/\sqrt{n}, \infty)$.

Clearly, we would reject H_0 whenever this interval shows no overlap

with the interval $(-\infty, \, \mu_0]$ as given by $H_0 : \mu \le \mu_0$.

Testing $H_0 : \mu \ge \mu_0$ vs. $H_1 : \mu < \mu_0$, reject when $\sqrt{n}(\bar{X} - \mu_0)/s \le t_{n-1,\alpha}$

or when the corresponding $100(1-\alpha)\%$ upper confidence bound interval

$(-\infty, \, \bar{X} - t_{n-1,\alpha} \times s/\sqrt{n})$ does not overlap the hypothesis interval $[\mu_0, \, \infty)$.

Return to Two-Sample Tests

Here we revisit the 2-sample problem, this time from a population perspective.

Assume $X_1, \ldots, X_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X, \sigma^2)$ is independent of $Y_1, \ldots, Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_Y, \sigma^2)$.

Note the assumptions of normality and equal variances (will be checked later).

The respective independence assumptions are more or less a judgment issue.

We want to test $H_0 : \mu_X = \mu_Y$ vs. $H_1 : \mu_X \neq \mu_Y$, with σ unknown and unspecified.

Clearly $\quad \bar{Y} - \bar{X} \sim \mathcal{N}\left(\mu_Y - \mu_X, \dfrac{\sigma^2}{n} + \dfrac{\sigma^2}{m}\right) \quad$ or under $H_0 \quad \dfrac{\bar{Y} - \bar{X}}{\sigma\sqrt{1/m + 1/n}} \sim \mathcal{N}(0,1)$

is a good indicator of $H_0 : \mu_Y - \mu_X = 0$ being true or not.

Under H_0 the sampling distribution of $\bar{Y} - \bar{X} \sim \mathcal{N}\left(0, \frac{\sigma^2}{n} + \frac{\sigma^2}{m}\right)$ is unknown because of the unknown σ.

Estimating σ^2

Have two estimates of σ^2: $s_X^2 = \frac{1}{m-1}\sum_{i=1}^{m}(X_i - \bar{X})^2$ and $s_Y^2 = \frac{1}{n-1}\sum_{j=1}^{n}(Y_j - \bar{Y})^2$

How to combine or pool them? $(s_X^2 + s_Y^2)/2$? or any other $\lambda s_X^2 + (1-\lambda)s_Y^2$?

$s_X^2 \sim \sigma^2\chi^2_{m-1}/(m-1) \implies \text{var}(s_X^2) = \sigma^4 \times 2(m-1)/(m-1)^2 = 2\sigma^4/(m-1)$.
Similarly $\text{var}(s_Y^2) = 2\sigma^4/(n-1)$ and independence of s_X^2 and s_Y^2 give us

$$\text{var}\left(\lambda s_X^2 + (1-\lambda)s_Y^2\right) = \lambda^2\,\frac{2\sigma^4}{m-1} + (1-\lambda)^2\,\frac{2\sigma^4}{n-1}$$

a quadratic in λ with clear minimum (calculus exercise) at $\lambda = (m-1)/(m+n-2)$.

This suggests $s^2 = \frac{(m-1)s_X^2}{m+n-2} + \frac{(n-1)s_Y^2}{m+n-2} = \frac{\sum_{i=1}^{m}(X_i - \bar{X})^2 + \sum_{j=1}^{n}(Y_j - \bar{Y})^2}{m+n-2}$

as “best” pooled variance estimate for σ^2.

$s^2 \sim \sigma^2\chi^2_{m+n-2}/(m+n-2)$ Is Independent of $\bar{Y}-\bar{X}$

$X_1,\ldots,X_m \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_X,\sigma^2) \quad \Longrightarrow \quad$ $\bar{X}$ and s_X^2 are independent.

$Y_1,\ldots,Y_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu_Y,\sigma^2) \quad \Longrightarrow \quad$ $\bar{Y}$ and s_Y^2 are independent.

$X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$ independent $\quad \Longrightarrow \quad$ $\bar{X}$, s_X^2 and $\bar{Y}$, s_Y^2 are all independent.

$\Longrightarrow \quad \bar{Y}-\bar{X}$ and $\displaystyle s^2 = \frac{(m-1)s_X^2}{m+n-2} + \frac{(n-1)s_Y^2}{m+n-2}$ are independent.

$(m+n-2)s^2 \sim \sigma^2\chi^2_{m+n-2}$ since $(m-1)s_X^2 \sim \sigma^2\chi^2_{m-1}$ and
$(n-1)s_Y^2 \sim \sigma^2\chi^2_{n-1}$, see slide 40 (sum of independent χ^2 random variables).

Two-Sample t-Statistic

Thus

$$t(\mathbf{X},\mathbf{Y}) = \frac{\bar{Y}-\bar{X}}{s\sqrt{1/m+1/n}} = \frac{(\bar{Y}-\bar{X})\Big/\left[\sigma\sqrt{1/m+1/n}\right]}{s/\sigma}$$

$$= \frac{Z}{\sqrt{C_{m+n-2}/(m+n-2)}} \sim t_{m+n-2}$$

gives us the desired 2-sample t-statistic with known null distribution under H_0.

Reject $H_0 : \mu_Y - \mu_X = 0$ at significance level α when

$$|t(\mathbf{X},\mathbf{Y})| \geq t_{m+n-2,1-\alpha/2} = t_{\text{crit}}\ .$$

The 2-sided p-value of the observed $t(\mathbf{x},\mathbf{y})$ is

$$P(|t_{m+n-2}| \geq |t(\mathbf{x},\mathbf{y}|) = 2 * (1 - \texttt{pt}(\texttt{abs}(\texttt{t}(\mathbf{x},\mathbf{y})),\texttt{m}+\texttt{n}-2))$$

Other Hypotheses

We may also want to test $H_\Delta : \mu_Y - \mu_X = \Delta$ or $H_\Delta : \mu_Y - \mu_X - \Delta = 0$

for a specified value Δ.

The natural change is to subtract Δ from $\bar{Y} - \bar{X}$ and note that

$$t(\mathbf{X}, \mathbf{Y} - \Delta) = \frac{\bar{Y} - \bar{X} - \Delta}{s\sqrt{1/m + 1/n}} = \frac{(\bar{Y} - \bar{X} - \Delta) \Big/ \left[\sigma\sqrt{1/m + 1/n}\right]}{s/\sigma} \sim t_{m+n-2}$$

when H_Δ is true.

We reject H_Δ whenever $|t(\mathbf{X}, \mathbf{Y} - \Delta)| \geq t_{m+n-2,1-\alpha/2}$ at significance level α.

As in the case of the 1-sample test we can derive (exercise)

a $100(1-\alpha)\%$ confidence interval for Δ as

$$\bar{Y} - \bar{X} \;\pm\; t_{m+n-2,1-\alpha/2} \times s \times \sqrt{1/m + 1/n}$$

Two-Sample `t.test` in R

```
> t.test(flux$SIR[flux$FLUX=="Y"],flux$SIR[flux$FLUX=="X"],var.equal=T)

        Two Sample t-test

data:  flux$SIR[flux$FLUX == "Y"] and flux$SIR[flux$FLUX == "X"]
t = -2.5122, df = 16, p-value = 0.0231
alternative hypothesis: true difference in means is not equal to 0
95 percent confidence interval:
 -2.7042872 -0.2290462
sample estimates:
mean of x mean of y
 9.133333 10.600000
```

Power Function of the Two-Sample t-Test

The power function of the 2-sided 2-sample t-test is given by

$$\begin{aligned}\beta(\mu_X,\mu_Y,\sigma) &= P(|t(\mathbf{X},\mathbf{Y})| \ge t_{\text{crit}}) \\ &= P(t(\mathbf{X},\mathbf{Y}) \le -t_{\text{crit}}) + 1 - P(t(\mathbf{X},\mathbf{Y}) < t_{\text{crit}})\end{aligned}$$

$$\begin{aligned}t(\mathbf{X},\mathbf{Y}) &= \frac{\bar{Y}-\bar{X}}{s\sqrt{1/m+1/n}} \\ &= \frac{\frac{\bar{Y}-\bar{X}-(\mu_Y-\mu_X)}{\sigma\sqrt{1/m+1/n}} + \frac{\mu_Y-\mu_X}{\sigma\sqrt{1/m+1/n}}}{s/\sigma} = \frac{Z+\delta}{\sqrt{\frac{C_{m+n-2}}{m+n-2}}} \sim t_{m+n-2,\delta}\end{aligned}$$

with noncentrality parameter $\delta = (\mu_Y - \mu_X)/[\sigma\sqrt{1/m+1/n}]$.

Thus β depends on μ_X,μ_Y,σ only through δ and we have

$$\beta(\delta) = \mathtt{pt}(-\mathtt{t}_{\mathtt{crit}},\mathtt{m}+\mathtt{n}-2,\delta) + 1 - \mathtt{pt}(\mathtt{t}_{\mathtt{crit}},\mathtt{m}+\mathtt{n}-2,\delta)$$

Allocation of Sample Sizes

Previously we saw that such power functions are increasing in $|\delta|$.

If we can allocate m and n subject to the restriction $m+n=N$ being fixed at some even number, we maximize $|\delta|$ for fixed $|\mu_Y - \mu_X|/\sigma$ by minimizing

$$\frac{1}{m}+\frac{1}{n}=\frac{1}{m}+\frac{1}{N-m}$$

over m.

Calculus or algebra $\implies$ $m=n=N/2$ gives us the minimum and thus the sample size allocation with highest power potential against any fixed $|\mu_Y - \mu_X|/\sigma$.

Sample Size Planning Tools

It is a simple matter to adapt the previous sample size functions `sample.size2` and `sample.size1` for the 1-sample t-test to the 2-sample situations, i.e., for 2-sided tests and 1-sided tests.

Since $t_{\text{crit}} = t_{m+n-2,1-\alpha/2}$ and $t_{m+n-2,\delta}$ (aside from δ) depend on m and n only through $N = m+n$ and since the previous slide made the case for $m = n = N/2$, we just need to express the corresponding sample size function in terms of N and use $\delta = (\mu_Y - \mu_x)/\left(\sigma\left(\sqrt{2/N+2/N}\right)\right) = \sqrt{N/4} \times (\mu_Y - \mu_x)/\sigma$.

Of course, we would then also need to take $m = n$.

See Homework!

Reflection on Treatments of Two-Sample Problem

Randomization test: No population assumptions. Under the hypothesis of no flux difference the SIR results for the 18 boards would be the same under all flux assignments. The flux assignments are then irrelevant!

Test is based on random assignment of fluxes giving us the randomization reference distribution for calculation of p-values or critical values.
Using $t(\mathbf{X}, \mathbf{Y})$, a good approximation to the null distribution often is t_{m+n-2}.

Generalization to other boards only by judgment or assumptions.

Normal 2-sample test: Assumes 2 independent samples from normal populations with common variance and possibly different means. We generalize upfront.

The t-test makes inferences concerning these two (conceptual) populations.

Check $\sigma_X^2 = \sigma_Y^2$

Test the hypothesis $H_0 : \sigma_X^2/\sigma_Y^2 = 1$ vs. the alternative $H_1 : \sigma_X^2/\sigma_Y^2 \neq 1$

A good indicator for σ_X^2/σ_Y^2 is the ratio of sample variances $F = s_X^2/s_Y^2$.

Note that

$$F = \frac{s_X^2}{s_Y^2} = \frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} \times \frac{\sigma_X^2}{\sigma_Y^2} \sim \frac{\sigma_X^2}{\sigma_Y^2} \times F_{m-1,n-1} \quad \Rightarrow \quad F = \frac{s_X^2}{s_Y^2} \sim F_{m-1,n-1} \quad \text{under} \quad H_0 \,.$$

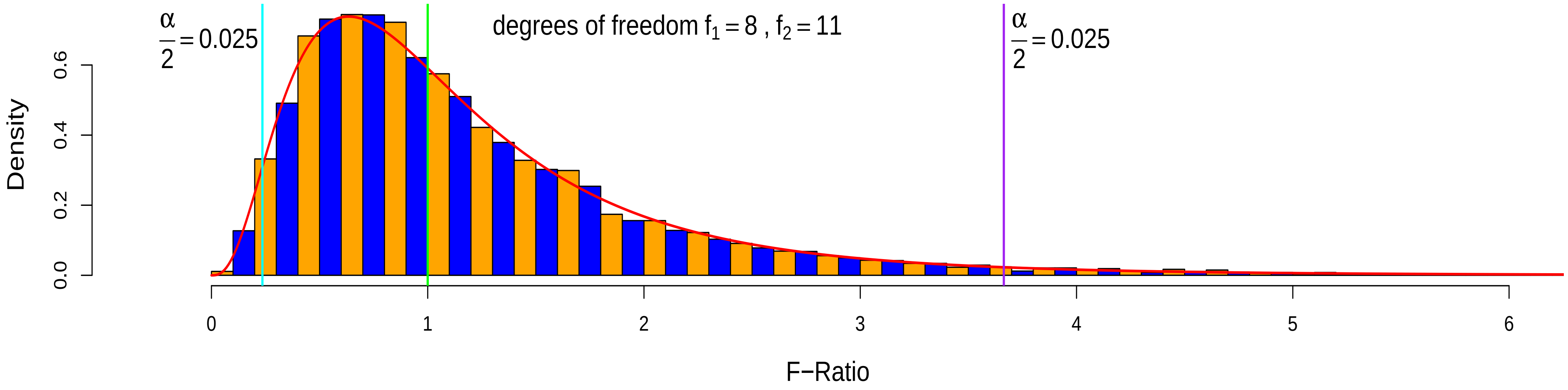
Thus reject H_0 when s_X^2/s_Y^2 is $\leq F_{m-1,n-1,\alpha/2}$ or $\geq F_{m-1,n-1,1-\alpha/2}$.

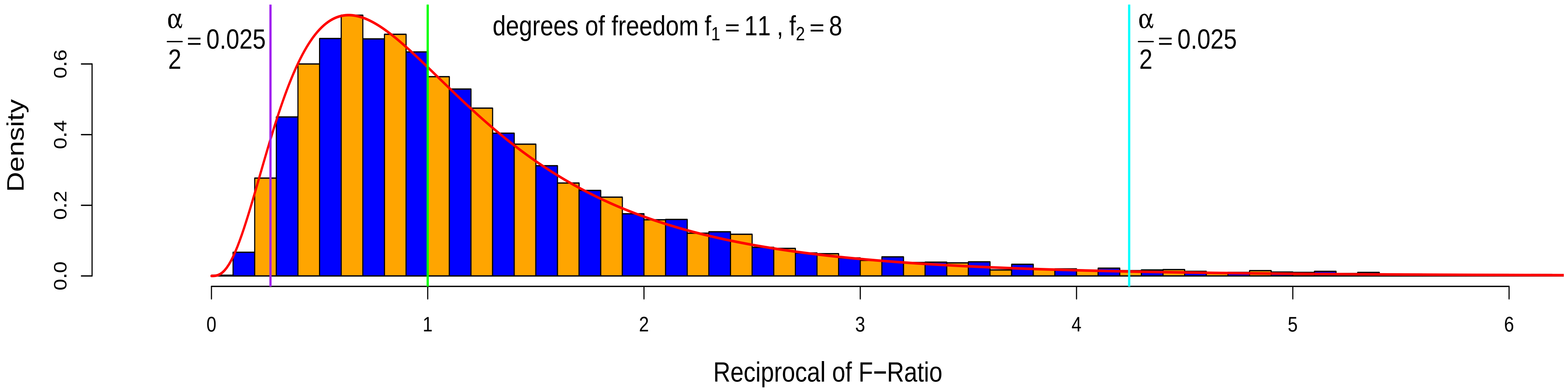
We denote by $F_{m-1,n-1,p}$ the p-quantile of $F_{m-1,n-1}$.

Unfortunately this test is very sensitive to deviations from normality

(will return to this later).

F-Distribution & Critical Values for $\alpha = .05$





Confidence Interval for σ_X^2/σ_Y^2

From

$$\frac{s_X^2/\sigma_X^2}{s_Y^2/\sigma_Y^2} = \frac{s_X^2/s_Y^2}{\sigma_X^2/\sigma_Y^2} \sim F_{m-1,n-1}$$

we get

$$\begin{aligned} 1-\alpha &= P\left(F_{m-1,n-1,\alpha/2} \leq \frac{s_X^2/s_Y^2}{\sigma_X^2/\sigma_Y^2} \leq F_{m-1,n-1,1-\alpha/2}\right) \\ &= P\left(\frac{s_X^2/s_Y^2}{F_{m-1,n-1,1-\alpha/2}} \leq \frac{\sigma_X^2}{\sigma_Y^2} \leq \frac{s_X^2/s_Y^2}{F_{m-1,n-1,\alpha/2}}\right) \end{aligned}$$

i.e.,

$$\left[(s_X^2/s_Y^2)/F_{m-1,n-1,1-\alpha/2},\ (s_X^2/s_Y^2)/F_{m-1,n-1,\alpha/2}\right]$$

is a $100(1-\alpha)\%$ confidence interval for σ_X^2/σ_Y^2.

$t^\star$-Test when $\sigma_X^2 \neq \sigma_Y^2$?

When $\sigma_X^2 \neq \sigma_Y^2$ we could emulate $(\bar{Y}-\bar{X})/\sqrt{\sigma_X^2/m+\sigma_Y^2/n} \sim \mathcal{N}(0,1)$

by using $t^\star(\mathbf{X},\mathbf{Y}) = (\bar{Y}-\bar{X})/\sqrt{s_X^2/m+s_Y^2/n}$ as test statistic.

But what is its reference distribution under $H_0 : \mu_X = \mu_Y$? It is unknown.

This is referred to as the Behrens-Fisher problem.

Approximate the distribution of $s_X^2/m+s_Y^2/n$ by that of $a \times \chi_f^2/f$, where a and f are chosen to match mean and variance of approximand and approximation.

$$E\left(a \times \chi_f^2/f\right) = a \qquad \text{and} \qquad \mathrm{var}\left(a \times \chi_f^2/f\right) = \frac{a^2 \times 2f}{f^2} = \frac{2a^2}{f}\ .$$

$$E\left(\frac{s_X^2}{m}+\frac{s_Y^2}{n}\right) = \frac{\sigma_X^2}{m}+\frac{\sigma_Y^2}{n} = a \quad \& \quad \mathrm{var}\left(\frac{s_X^2}{m}+\frac{s_Y^2}{n}\right) = \frac{2(m-1)\sigma_X^4}{m^2(m-1)^2}+\frac{2(n-1)\sigma_Y^4}{n^2(n-1)^2} = \frac{2a^2}{f}$$

The Satterthwaite Approximation

$$\Longrightarrow \quad f = \frac{a^2}{\frac{\sigma_X^4}{m^2(m-1)} + \frac{\sigma_Y^4}{n^2(n-1)}} = \frac{\left(\sigma_X^2/m + \sigma_Y^2/n\right)^2}{\frac{(\sigma_X^2/m)^2}{m-1} + \frac{(\sigma_Y^2/n)^2}{n-1}}$$

Replace the unknown σ_X^2 and σ_Y^2 by s_X^2 and s_Y^2 and use instead $\hat{f}$, where

$$\hat{f} = \frac{\left(s_X^2/m + s_Y^2/n\right)^2}{\frac{(s_X^2/m)^2}{m-1} + \frac{(s_Y^2/n)^2}{n-1}} \quad \text{and approximate} \quad \frac{s_X^2/m + s_Y^2/n}{\sigma_X^2/m + \sigma_Y^2/n} \approx \chi^2_{\hat{f}}/\hat{f}$$

$$\Longrightarrow \quad t^\star(\mathbf{X},\mathbf{Y}) = \frac{(\bar{Y} - \bar{X})/\sqrt{\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}}}{\sqrt{\left(\frac{s_X^2}{m} + \frac{s_Y^2}{n}\right) \Big/ \left(\frac{\sigma_X^2}{m} + \frac{\sigma_Y^2}{n}\right)}} = \frac{Z}{\sqrt{\chi^2_{\hat{f}}/\hat{f}}} \approx t_{\hat{f}}$$

Reject $H_0 : \mu_X = \mu_Y$ when $|t^\star(\mathbf{X},\mathbf{Y})|$ is too large and

compute p-values from the $t_{\hat{f}}$ distribution.

Which Test to Use: t or $t^\star$?

When $m = n$ one easily sees that $t(\mathbf{X}, \mathbf{Y}) = t^\star(\mathbf{X}, \mathbf{Y})$, but their null distributions are only the same when $s_X^2 = s_Y^2$ in which case $\hat{f} = 2(m-1)$.
However, $s_X^2 = s_Y^2$ is an unlikely occurrence. To show only takes some algebra.

What happens when $m = n$ but $\sigma_X \neq \sigma_Y$ and we use $t(\mathbf{X}, \mathbf{Y})$ anyway?

What happens when $m \neq n$ and $\sigma_X \neq \sigma_Y$ and we use $t(\mathbf{X}, \mathbf{Y})$ anyway?

How is the probability of type I error affected?

Such questions can easily be examined using simulation in R.

When $m = n$ or $\sigma_X \approx \sigma_Y$ it seems that using $t(\mathbf{X}, \mathbf{Y})$is relatively safe, otherwise use $t^\star(\mathbf{X}, \mathbf{Y})$ (see following slides).

Simulating the Null-Distribution of $t(\mathbf{X},\mathbf{Y})$

Recall

$$t(\mathbf{X},\mathbf{Y}) = \frac{\bar{Y}-\bar{X}}{s\sqrt{1/n+1/m}} \qquad \text{with} \qquad s^2 = \frac{(m-1)s_X^2+(n-1)s_Y^2}{m+n-2}$$

and under $H_0 : \mu_X = \mu_Y$ with independent

$$\bar{Y}-\bar{X} \sim \mathcal{N}(0, \sigma_Y^2/n+\sigma_X^2/m)\,, \quad (m-1)s_X^2 \sim \sigma_X^2\chi_{m-1}^2 \quad \text{and} \quad (n-1)s_Y^2 \sim \sigma_Y^2\chi_{n-1}^2$$

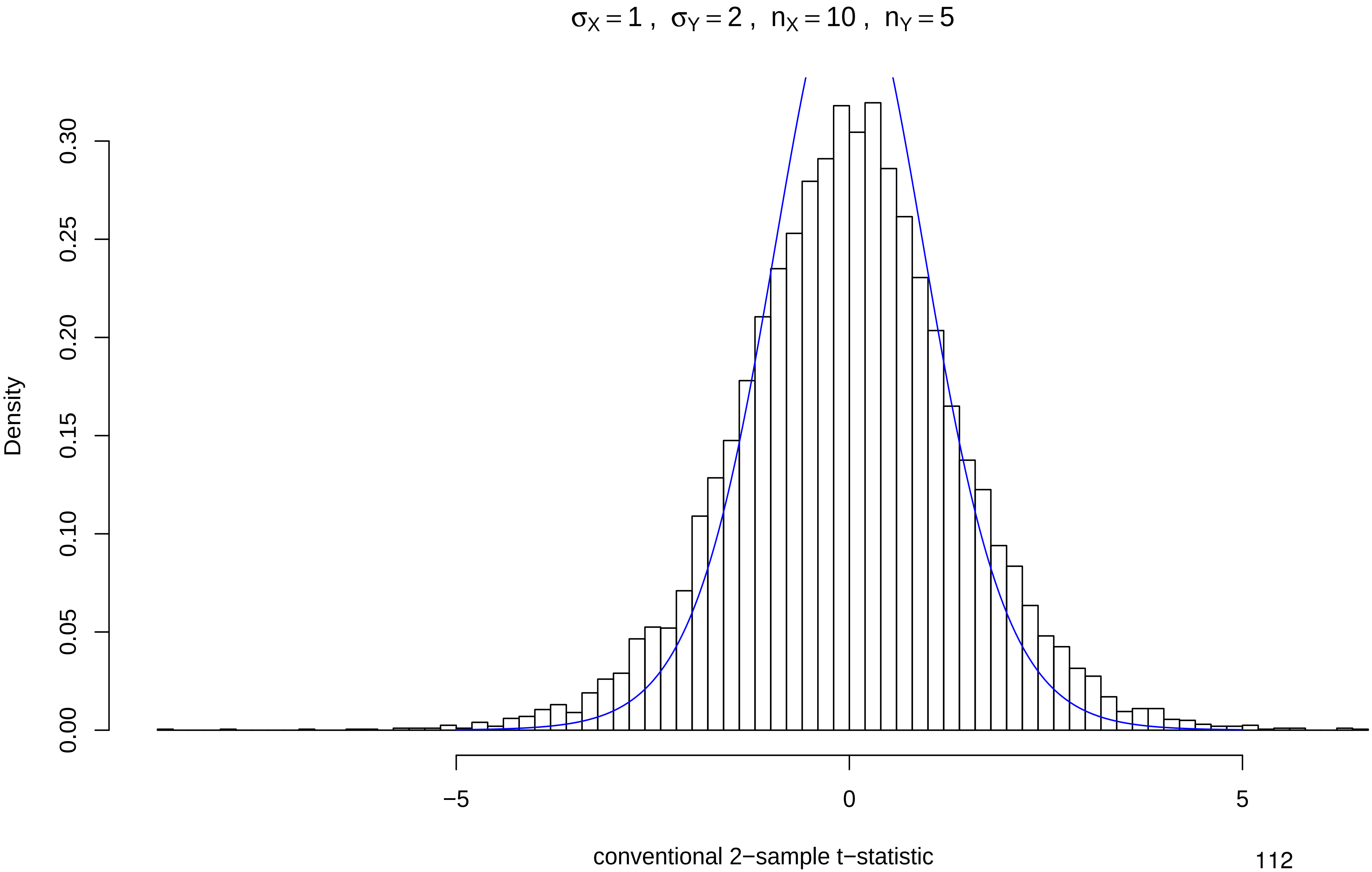
This leads to the first 3 command lines in the R function `t.sig.diff`, i.e.:

```
Dbar=rnorm(Nsim,0,sqrt(sigX^2/m+sigY^2/n))
s2=(rchisq(Nsim,m-1)*sigX^2+rchisq(Nsim,n-1)*sigY^2)/(m+n-2)
t.stat=Dbar/sqrt(s2*(1/m+1/n))
```

R Function Examining P(Type I Error) for $t(\mathbf{X},\mathbf{Y})$-Test

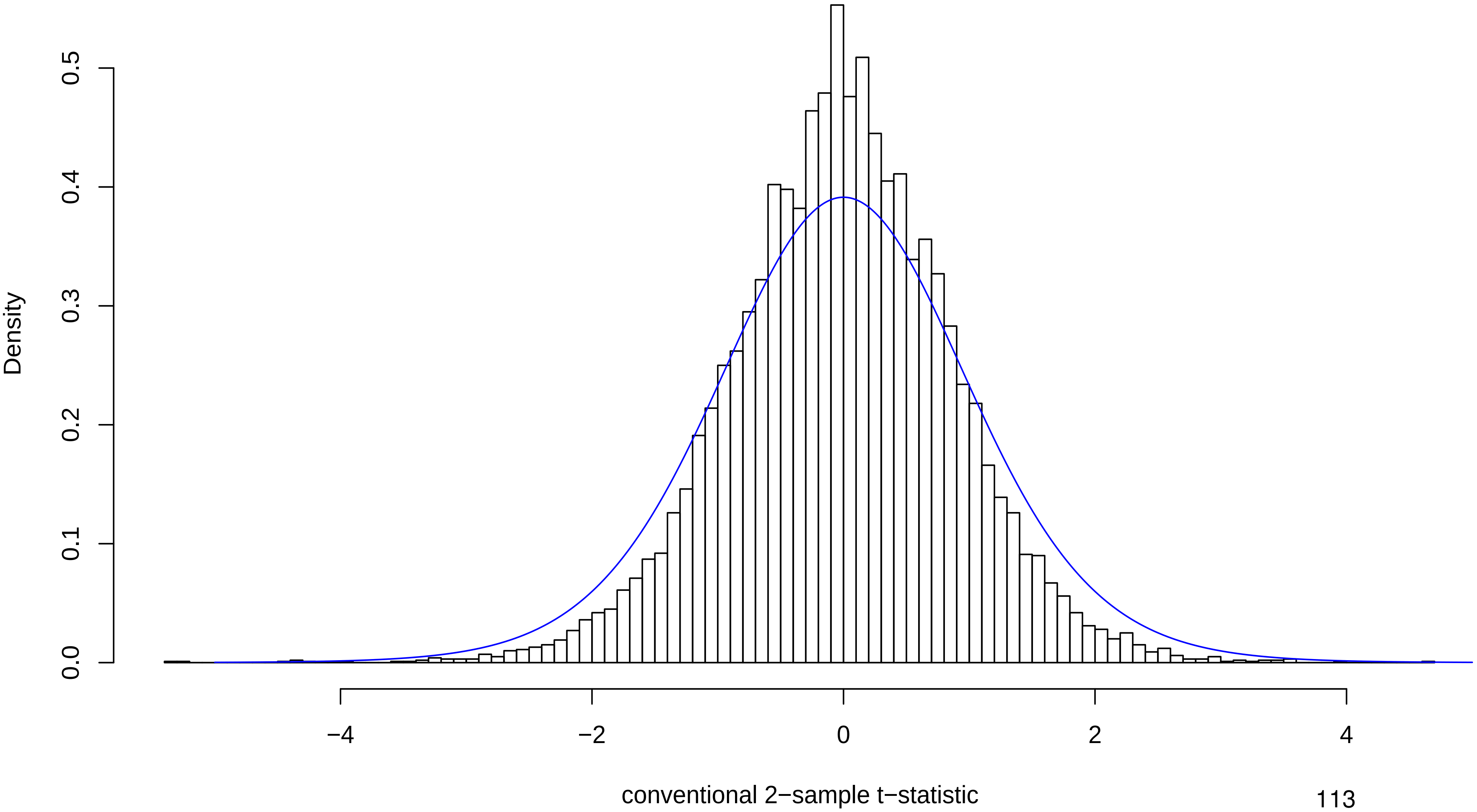
```
t.sig.diff = function (m=10,n=5,sigX=1,sigY=1,Nsim=10000)
{
Dbar=rnorm(Nsim,0,sqrt(sigX^2/m+sigY^2/n))
s2=(rchisq(Nsim,m-1)*sigX^2+rchisq(Nsim,n-1)*sigY^2)/(m+n-2)
t.stat=Dbar/sqrt(s2*(1/m+1/n))
x=seq(-5,5,.01)
y=dt(x,m+n-2)
hist(t.stat,nclass=101,probability=T,main="",
xlab="conventional 2-sample t-statistic")
title(substitute(sigma[X]==sigX~", "~
sigma[Y]==sigY~", "~n[X]==nX~", "~n[Y]==nY,
list(sigX=sigX,sigY=sigY,nX=m,nY=n)))

lines(x,y,col="blue")
}
```

Inflated P(Type I Error)



Deflated P(Type I Error)

$\sigma_X = 2\ ,\ \ \sigma_Y = 1\ ,\ \ n_X = 10\ ,\ \ n_Y = 5$



Heuristic Explanation of Opposite Effects

Recall that

$$s^2 = \frac{(n_X - 1)s_X^2 + (n_Y - 1)s_Y^2}{n_X + n_Y - 2} = \frac{9}{13}s_X^2 + \frac{4}{13}s_Y^2$$

Thus when $\sigma_X = 1$ and $\sigma_Y = 2$ we have

$$E(s^2) = \frac{9}{13}\sigma_X^2 + \frac{4}{13}\sigma_Y^2 = \frac{9}{13}1^2 + \frac{4}{13}2^2 = \frac{25}{13} = 1.923$$

and when $\sigma_X = 2$ and $\sigma_Y = 1$ we have

$$E(s^2) = \frac{9}{13}\sigma_X^2 + \frac{4}{13}\sigma_Y^2 = \frac{9}{13}2^2 + \frac{4}{13}1^2 = \frac{40}{13} = 3.077$$

while for $\sigma_X = \sigma_Y = 1.5$ we have (correct t-distribution)

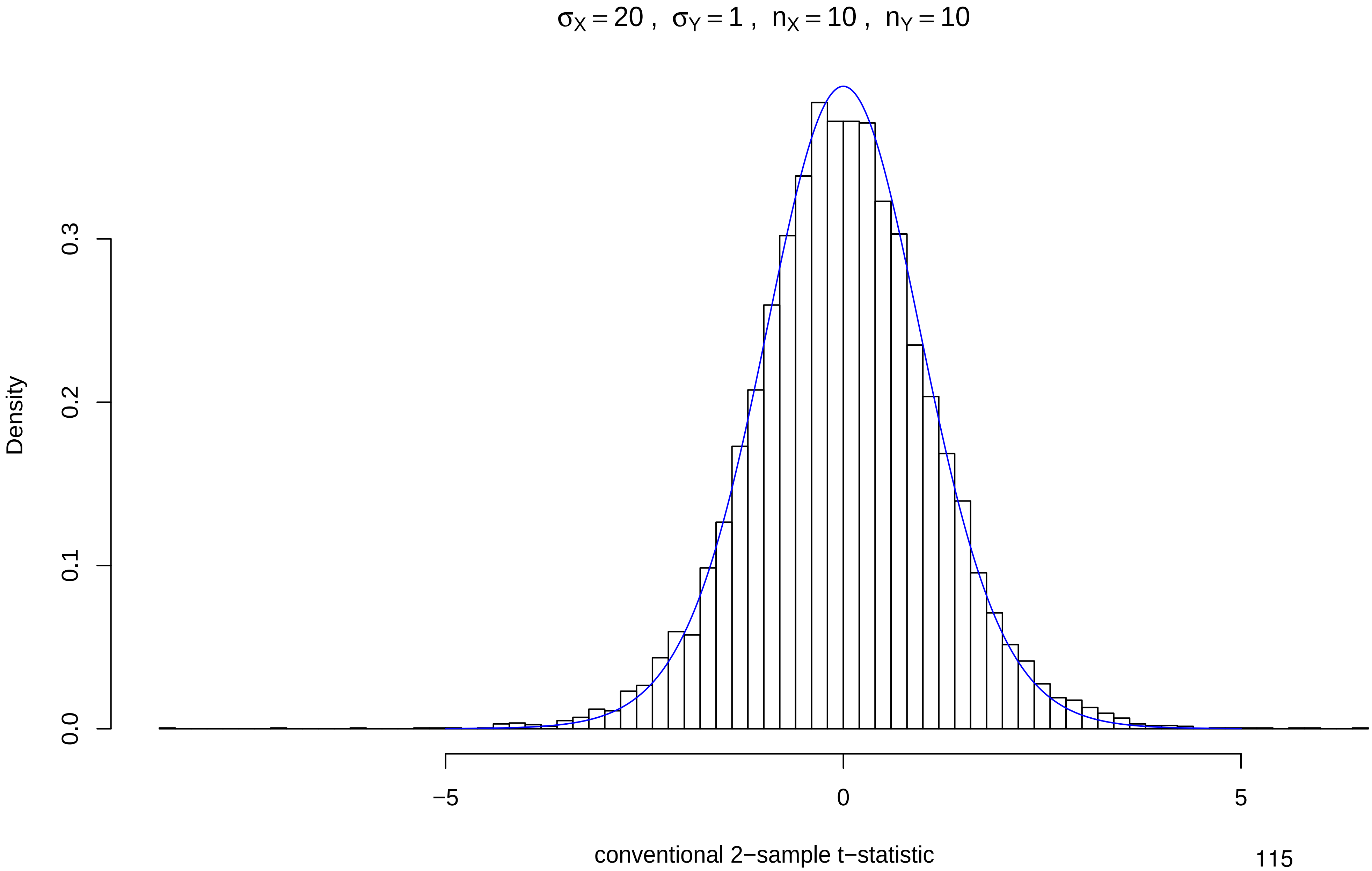
$$E(s^2) = \frac{9}{13}\sigma_X^2 + \frac{4}{13}\sigma_Y^2 = \frac{9}{13}1.5^2 + \frac{4}{13}1.5^2 = 2.25$$

We can clearly link the effect of $E(s^2)$ on the simulated distributions.

s^2 tends to be too small when $\sigma_X = 1, \sigma_Y = 2$ and too large when $\sigma_X = 2, \sigma_Y = 2$,

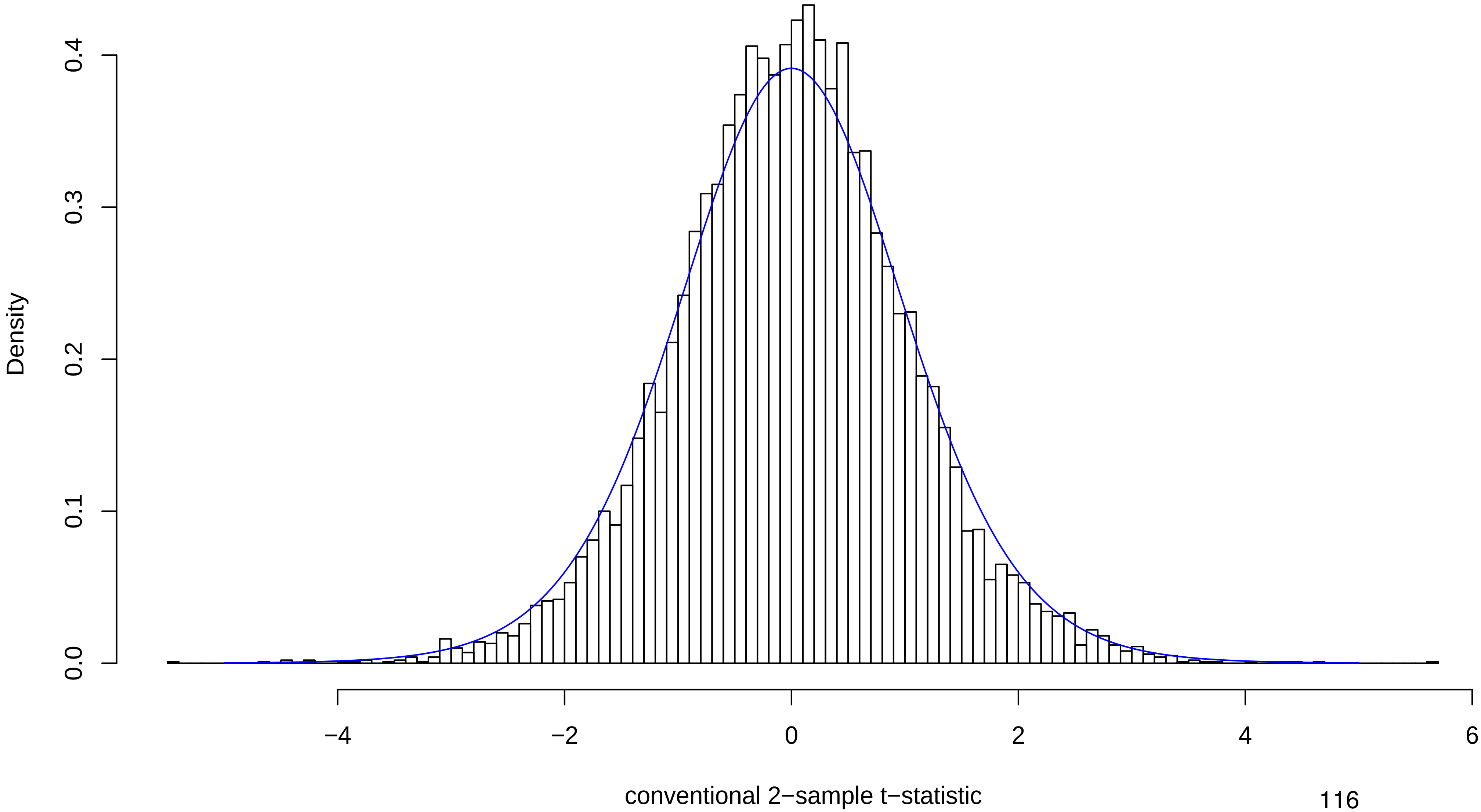
leading to inflated (deflated) values of $|t(\mathbf{X}, \mathbf{Y})|$, respectively.

P(Type I Error) Hardly Affected



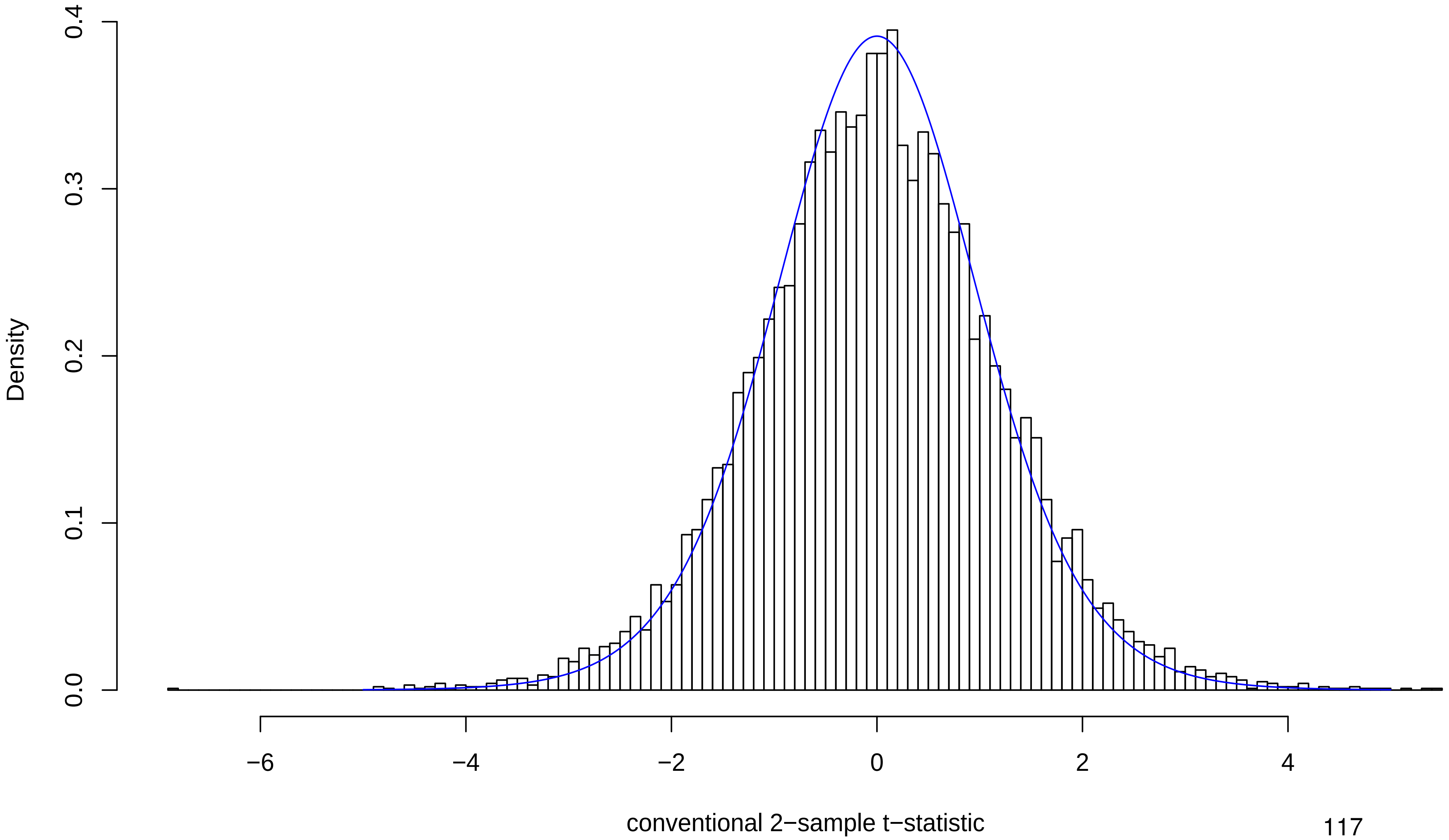
P(Type I Error) Mildly Affected

$\sigma_X = 1.2\ ,\ \ \sigma_Y = 1\ ,\ \ n_X = 10\ ,\ \ n_Y = 5$



P(Type I Error) Mildly Affected

$\sigma_X = 1\ ,\ \ \sigma_Y = 1.2\ ,\ \ n_X = 10\ ,\ \ n_Y = 5$



Checking Normality of a Sample

The p-quantile of $\mathcal{N}(\mu,\sigma^2)$ is $x_p = \mu + \sigma z_p$, z_p is the standard normal p-quantile.

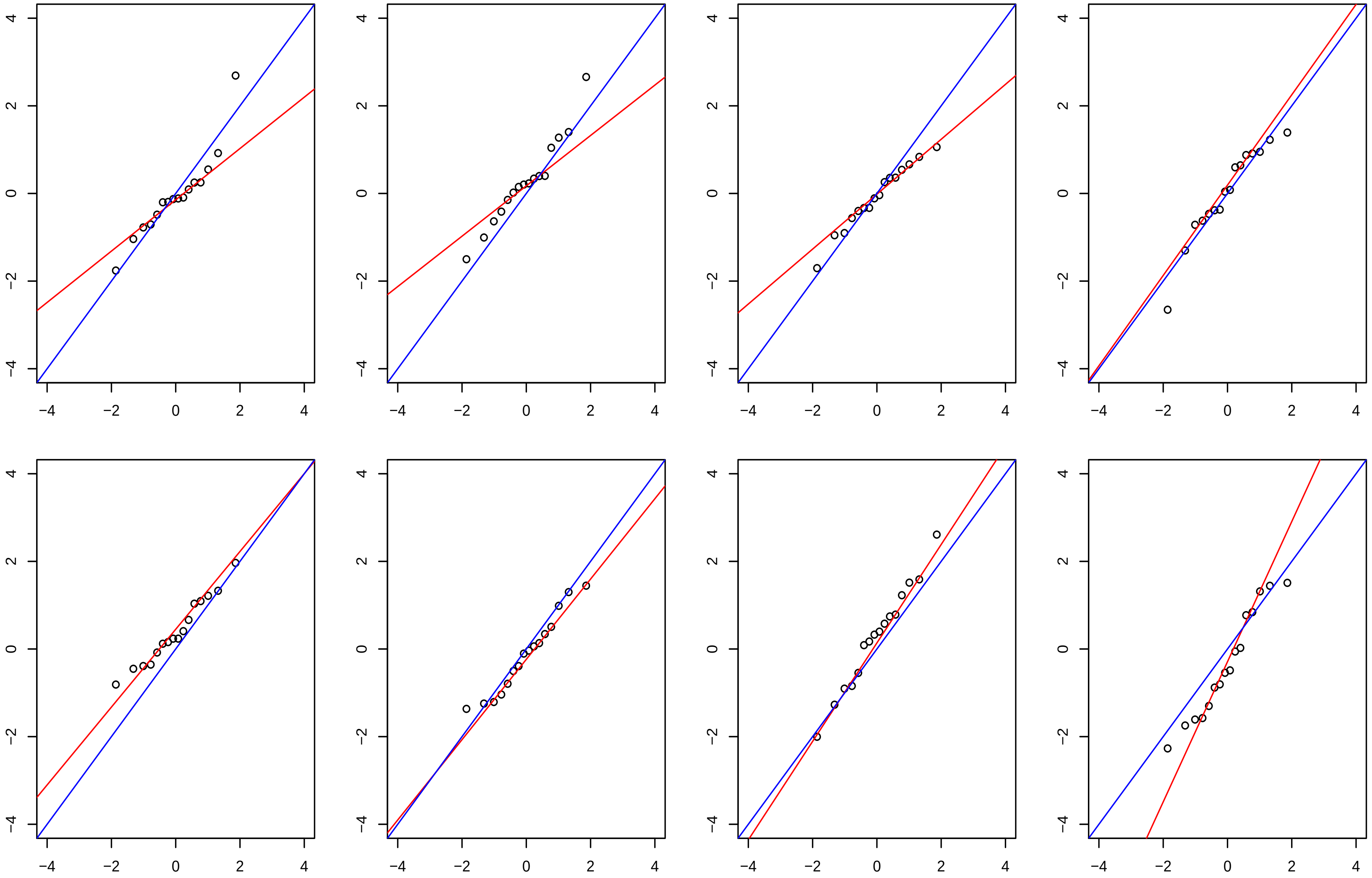
Sort the sample $X_1,\ldots,X_n$ in increasing order $X_{(1)} \le \ldots \le X_{(n)}$ assigning fractional ranks $p_i \in (0,1)$ to these order statistics in one of several ways for $i = 1,\ldots,n$:

$$p_i = \frac{i-.5}{n} \qquad \text{or} \qquad p_i = \frac{i}{n+1} \qquad \text{or} \qquad p_i = \frac{i-.375}{n+.25}\ .$$

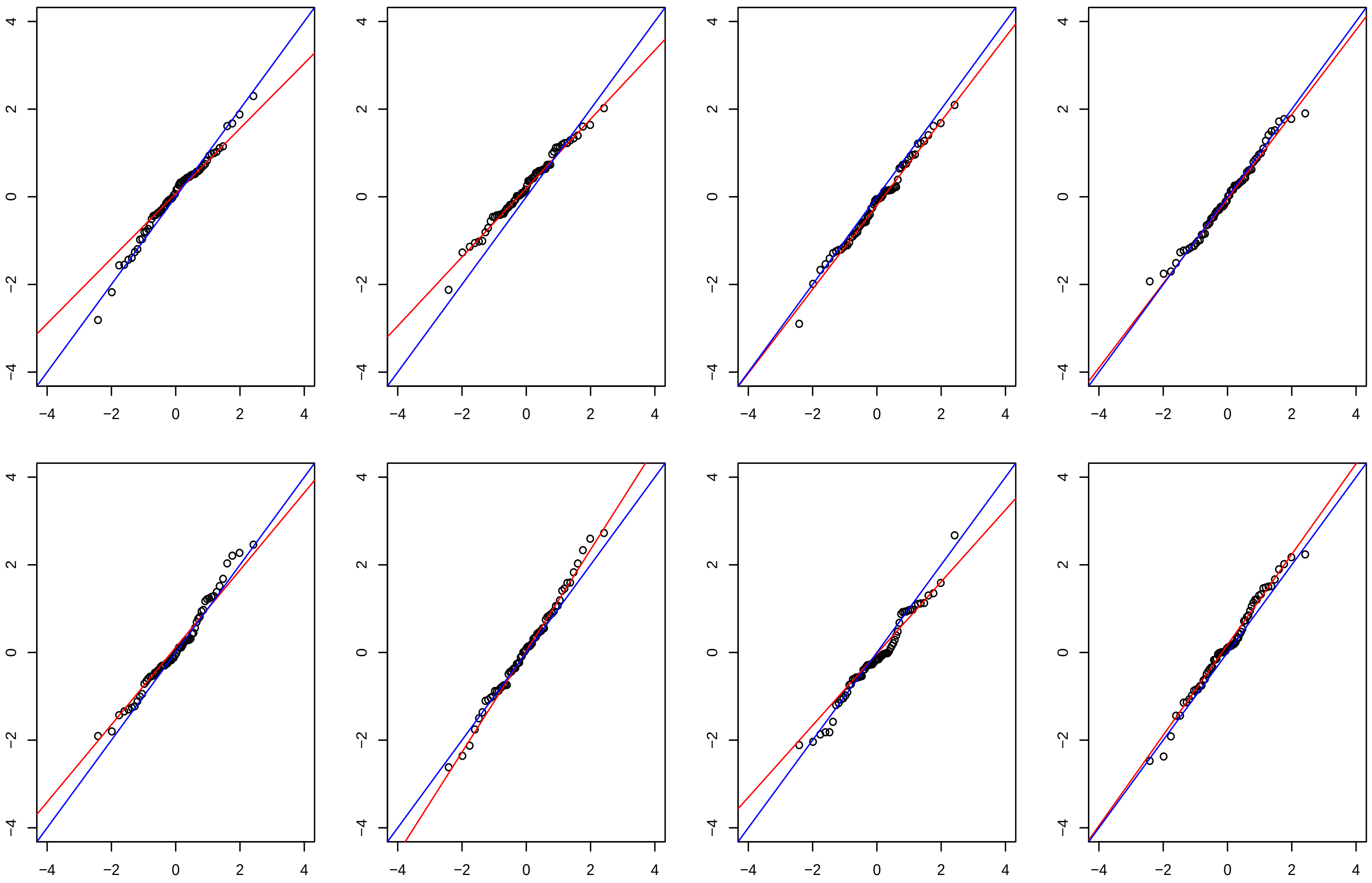
Plot $X_{(i)}$ against the standard normal p_i-quantile $z_{p_i} = \mathtt{qnorm}(\mathtt{p_i})$ for $i = 1,\ldots,n$. We would expect $X_{(i)} \approx x_{p_i} = \mu + \sigma z_{p_i}$, i.e., $X_{(i)}$ should look $\approx$ linear against z_{p_i} with intercept $\approx \mu$ and slope $\approx \sigma$. Judging approximate linearity takes practice.

The third choice for p_i is used by R in `qqnorm(x)` for a given sample vector `x`. `qqline(x)` fits a line to the middle half of the data.

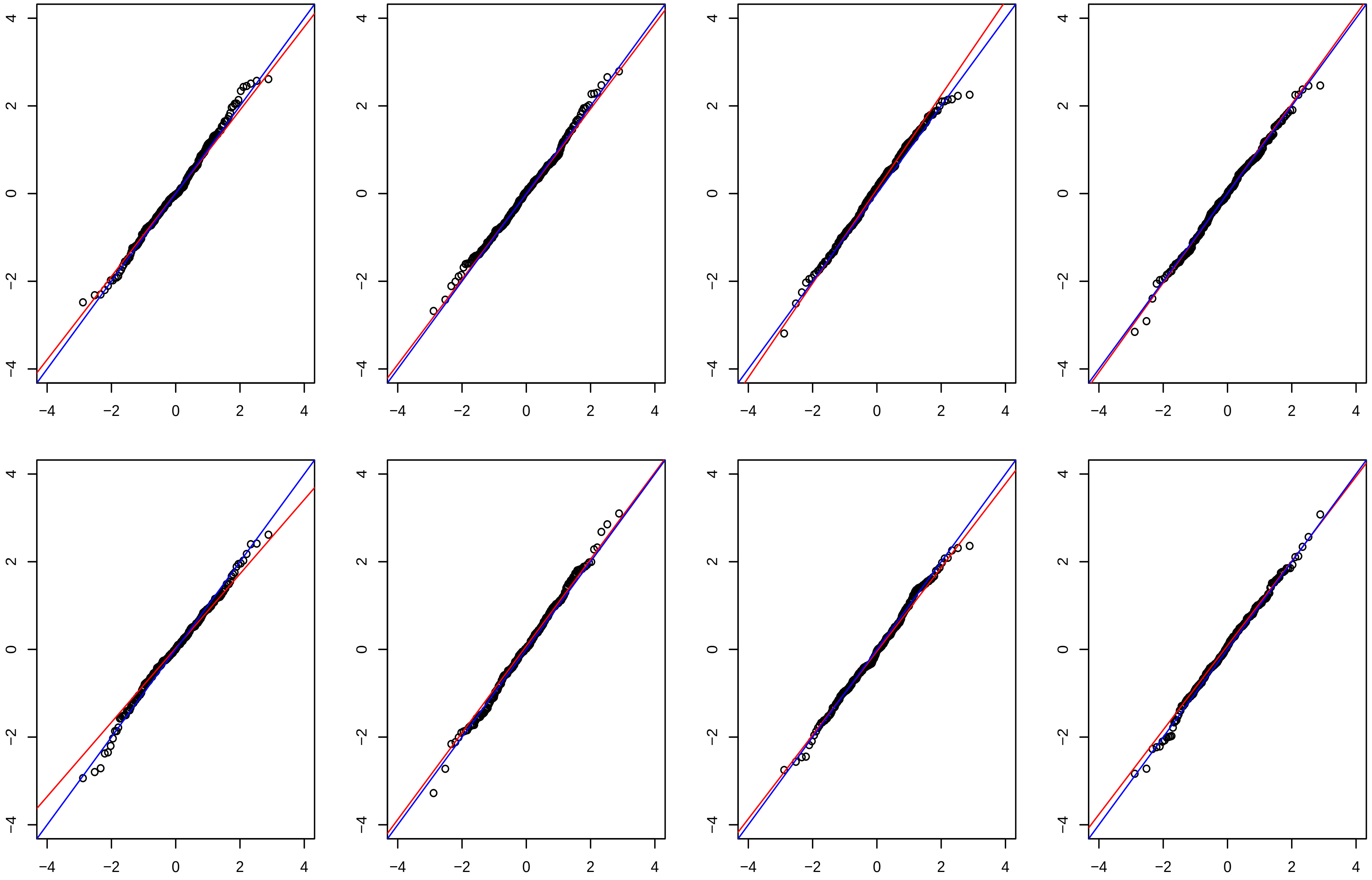
Normal QQ-Plot: $n = 16$



Normal QQ-Plot: $n = 64$



Normal QQ-Plot: $n = 256$



EDF-Based Tests of Fit

Judgment?? We can also carry out formal EDF-based tests of fit for normality.

Assume $X_1, \ldots, X_n \sim G$. Test $H_0 : G(x) = \Phi((x-\mu)/\sigma)$ for some μ and σ with μ and σ unspecified and unknown (a composite hypothesis).

The empirical distribution function (EDF) $\hat{F}_n(x)$ is defined as

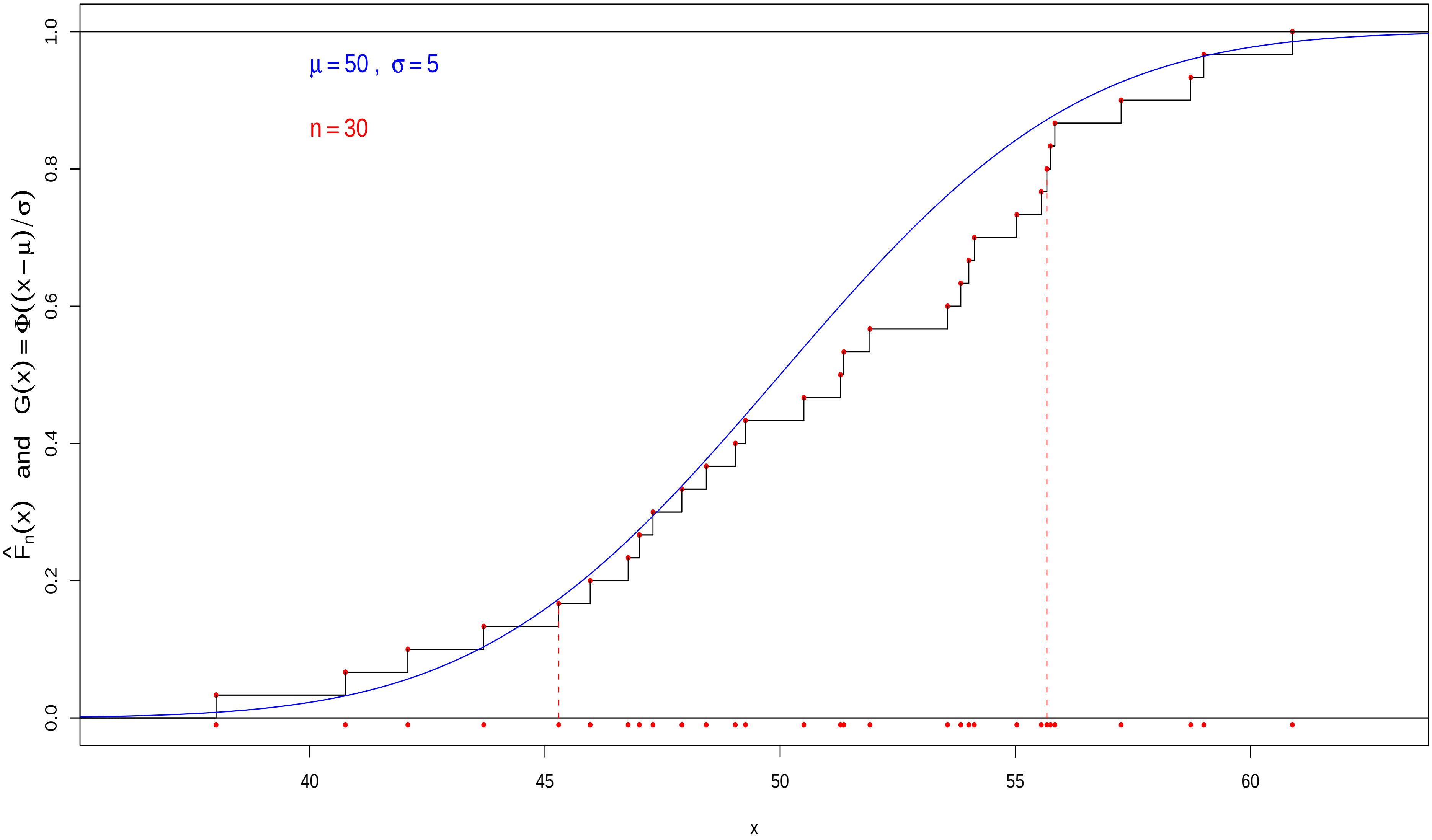
$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I_{(-\infty,\, x]}(X_i)$ with $B_i(x) = I_{(-\infty,\, x]}(X_i) = 1$ or 0 as $X_i \leq x$ or $x < X_i$.

$\hat{F}_n(x) =$ proportion of sample values $X_1, \ldots, X_n$ that are $\leq x$ ($\equiv$ success).

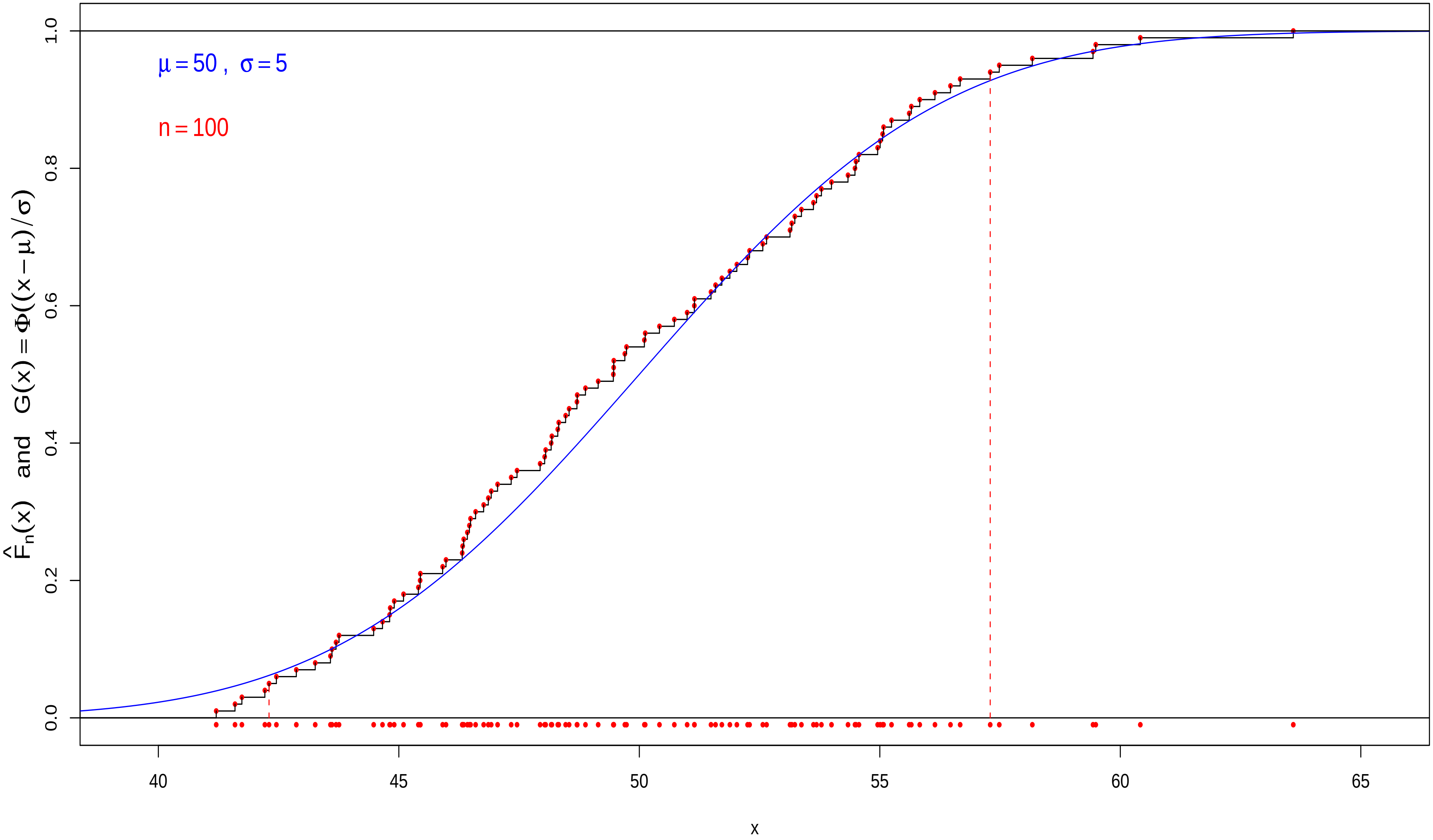
Here $B_1(x), \ldots, B_n(x)$ is an i.i.d. sequence of Bernoulli random variables with success probability $p = p(x) = P(X_i \leq x) = G(x)$.

Law of Large Numbers (LLN) $\Longrightarrow$ $\bar{B}(x) = \hat{F}_n(x) \longrightarrow G(x)$ as $n \to \infty$, for all x.

Empirical CDF for $n = 30$



Empirical CDF for $n = 100$



Compare the EDF with What?

The previous two slides compared the EDF with the CDF $G(x) = \Phi((x-\mu)/\sigma)$ from which the data were sampled.

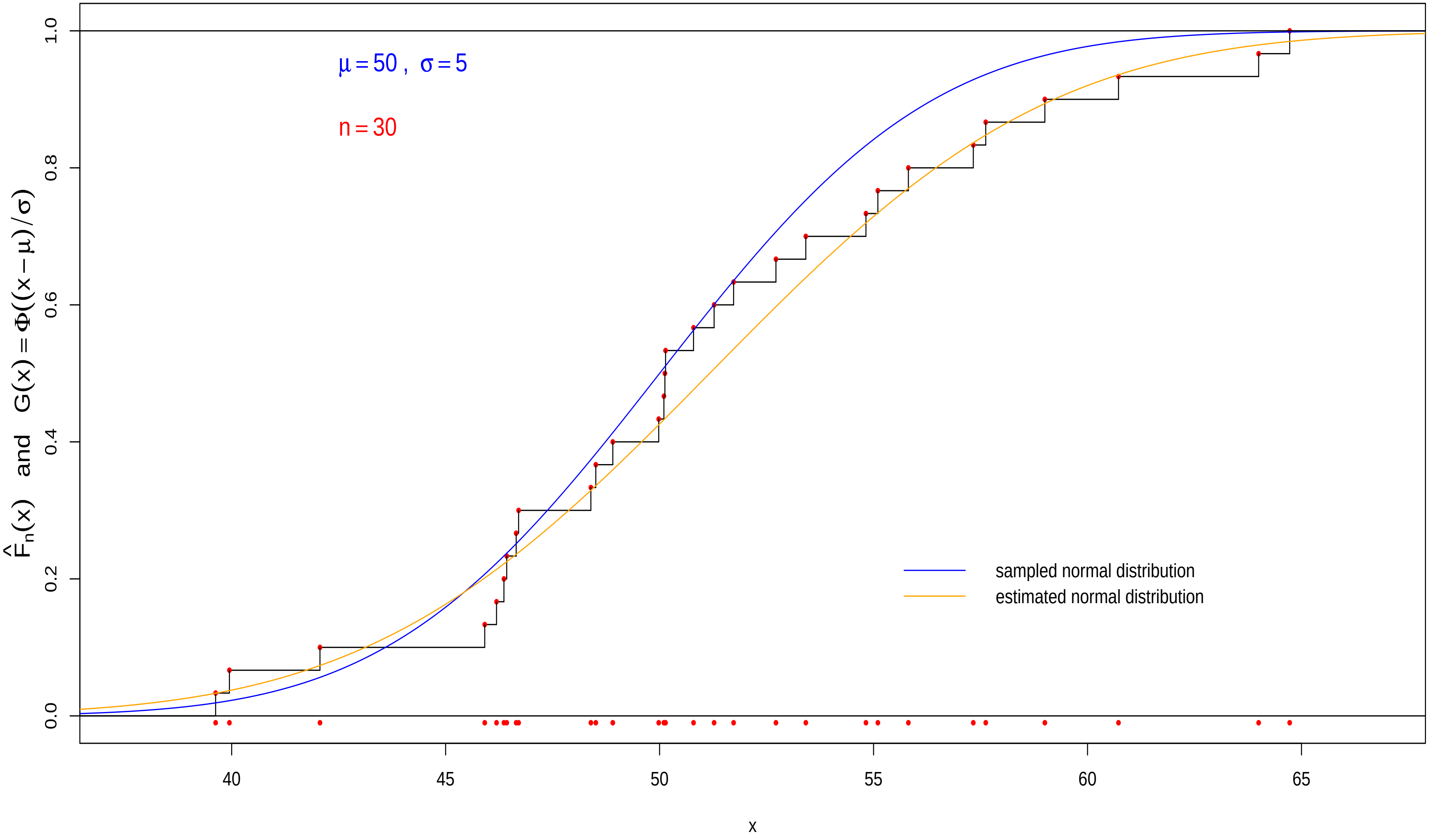
This was done to illustrate the validity of the LLN. However, we don't know μ and σ.

We can estimate $G(x)$ using $\hat{G}_n(x) = \Phi((x-\bar{X})/s) \quad \approx \quad G(x)$ for large n. This approximation is reasonable not only for normal samples but also as long as $\bar{X} \longrightarrow \mu$ and $s \longrightarrow \sigma$ for large n.

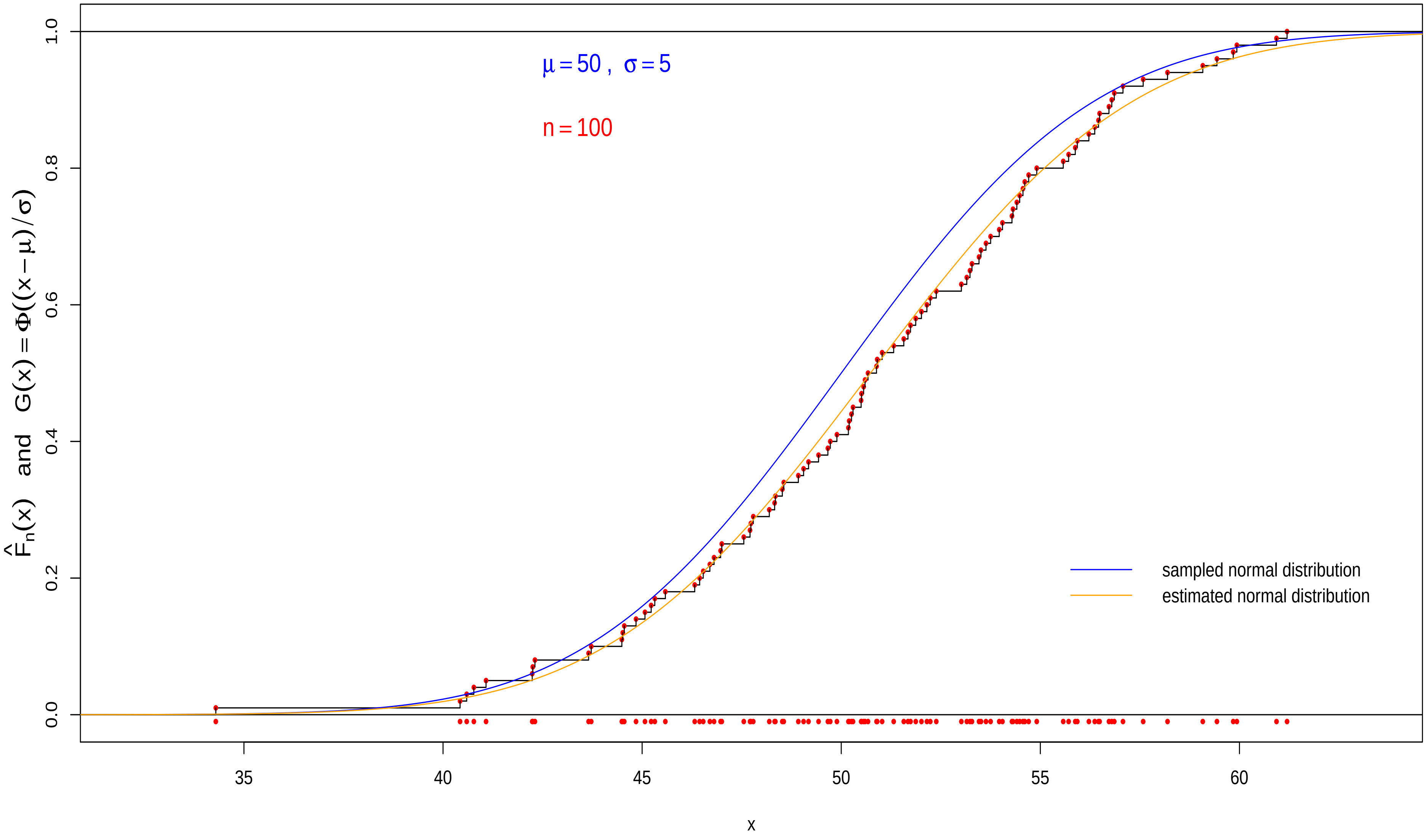
Compare $\hat{F}_n(x)$ with $\hat{G}_n(x)$ as proxy for comparing $\hat{F}_n(x)$ with $G(x)$. Compare them via some discrepancy metric $D(\hat{F}_n, \hat{G}_n)$.

We reject H_0 whenever $D(\hat{F}_n, \hat{G}_n)$ is too large, using the null distribution of $D(\hat{F}_n, \hat{G}_n)$ to find critical values or p-values.

Empirical CDF for $n = 30$ with Estimated CDF



Empirical CDF for $n = 100$ with Estimated CDF



Some Discrepancy Metrics

Note the generally closer fit $\hat{G}_n(x) \approx \hat{F}_n(x)$, as compared to $G(x) \approx \hat{F}_n(x)$.

$\hat{F}_n(x)$ represents the sample and $\hat{G}_n(x)$ is fitted to the sample.

$$D_{KS}(\hat{F}_n, \hat{G}_n) = \sup_x |\hat{F}_n(x) - \hat{G}_n(x)|$$

Kolmogorov-Smirnov criterion

$$D_{CvM}(\hat{F}_n, \hat{G}_n) = n \int_{-\infty}^{\infty} \left(\hat{F}_n(x) - \hat{G}_n(x)\right)^2 \hat{g}_n(x)\, dx$$

Cramér-von Mises criterion

$$D_{AD}(\hat{F}_n, \hat{G}_n) = n \int_{-\infty}^{\infty} \frac{\left(\hat{F}_n(x) - \hat{G}_n(x)\right)^2}{\hat{G}_n(x)(1 - \hat{G}_n(x))} \hat{g}_n(x)\, dx$$

Anderson-Darling criterion

Here $\hat{g}_n(x)$ is the density of $\hat{G}_n(x)$, i.e., $\hat{g}_n(x) = \varphi((x - \bar{X})/s)/s$,

where $\varphi(z)$ is the standard normal density.

Interpretation of Discrepancy Metrics

D_{KS} captures the local maximum discrepancy (at some x) between $\hat{G}_n(x)$ and $\hat{F}_n$.

D_{CvM} captures an accumulated (integrated) squared and weighted discrepancy between $\hat{G}_n(x)$ and $\hat{F}_n$, weighting via $\hat{g}_n(x)$.

D_{AD} captures an accumulated (integrated) squared discrepancy between $\hat{G}_n(x)$ and $\hat{F}_n$ with especially high weights in the distribution tails in addition to $\hat{g}_n(x)$, i.e., when $\hat{G}_n(x)$ or $1 - \hat{G}_n(x)$ are small.

The squaring of the discrepancies $\left(\hat{F}_n(x) - \hat{G}_n(x)\right)^2$ was done mainly for mathematical ease.

The absolute discrepancy $|\hat{F}_n(x) - \hat{G}_n(x)|$ is not so easily dealt with.

Computational Formulas for the Discrepancy Metrics

$$D_{KS}(\hat{F}_n,\hat{G}_n) = \max\left\{\max_i\left[i/n-\hat{G}_n(X_{(i)})\right],\ \max_i\left[\hat{G}_n(X_{(i)})-(i-1)/n\right]\right\}$$

$$D_{CvM}(\hat{F}_n,\hat{G}_n) = \sum_{i=1}^{n}\left[\hat{G}_n(X_{(i)})-(2i-1)/(2n)\right]^2+1/(12n)$$

$$D_{AD}(\hat{F}_n,\hat{G}_n) = -n-(1/n)\sum_{i=1}^{n}(2i-1)\left[\log(\hat{G}_n(X_{(i)}))+\log(1-\hat{G}_n(X_{(i)}))\right]$$

Here $X_{(1)} \le X_{(2)} \le \ldots \le X_{(n)}$ are the order statistics of the sample $X_1,\ldots,X_n$, i.e., its values in increasing order.

Note that the distribution of $\hat{G}_n(X_{(i)}) = \Phi((X_{(i)}-\bar{X})/s)$ does not depend on the unknown parameters μ and σ since

$$\frac{X_{(i)}-\bar{X}}{s} = \frac{(X_{(i)}-\mu)/\sigma-(\bar{X}-\mu)/\sigma}{s/\sigma} = \frac{Z_{(i)}-\bar{Z}}{s_Z} \quad \text{with} \quad Z_1,\ldots,Z_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1).$$

Null Distributions for D_{KS}, D_{CvM}, and D_{AD}

Approximate null distributions have been developed for all three metrics.

These are based on limiting null distributions ($n \longrightarrow \infty$) and substantial simulations for small and moderate sample sizes.

Goodness-of-Fit Techniques, (1986) ed. by R.B. D'Agostino and M.A. Stephens
See this reference for tabulations of critical values.

With today's computing power it is relatively easy to simulate p-values for observed discrepancy metrics.

One could even try other metrics for which limiting distribution results have not yet been investigated because of their analytical intractability.

The Package `nortest`

Fortunately the package `nortest` provides functions that evaluate each of the three discrepancy metrics and their corresponding p-values.

Install the package `nortest` directly from the web or from the zip file `nortest_1.0.zip` (available on my class web site) in your working directory. Do this installation just once for each R installation.

Invoke `library(nortest)` for each R session during which you want to use it.

The package `nortest` contains the routines:

`lillie.test` (D_{KS}), `cvm.test` (D_{CvM}), and `ad.test` (D_{AD})

See documentation: `?lillie.test`, `?cvm.test`, and `?ad.test`.

Kolmogorov-Smirnov Test for Normality

```
> lillie.test(rnorm(7)) # testing a normal sample
                        # of size n=7 for normality

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  rnorm(7)
D = 0.287, p-value = 0.08424

> lillie.test(runif(137)) # testing a uniform sample
                          # of size n=137 for normality

        Lilliefors (Kolmogorov-Smirnov) normality test

data:  runif(137)
D = 0.0877, p-value = 0.01169
```

Anderson-Darling Test for Normality

```
> ad.test(rnorm(10)) # testing a normal sample
                     # of size n=10 for normality

        Anderson-Darling normality test

data:  rnorm(10)
A = 0.4216, p-value = 0.2572

> ad.test(runif(30))   # testing a uniform sample
                       # of size n=30 for normality

        Anderson-Darling normality test

data:  runif(30)
A = 0.8551, p-value = 0.02452
```

General Comments on Goodness-of-Fit (GOF) Tests

Denote by H_0 our distributional hypothesis (here normality).

The following comments apply equally well to other distributional hypotheses.

For small sample sizes GOF tests tend to be very forgiving.

We reject only for gross deviations from H_0.

Very large samples from real applications most often lead to rejection of H_0.

The reason is that such tests are all consistent, i.e., they will reject H_0 for any alternative to H_0, provided the sample is large enough.

Such alternatives may look very similar to H_0, but not exactly the same.

Should we be concerned about such rejections?

Are tiny deviations from normality relevant? The “curse” of large n?!

GOF tests are most useful for moderate and not too large sample sizes.

A Simulation Experience

A client wanted to extrapolate very costly simulation results from $N_{\text{sim}} = 2000$ simulations far out into the distribution tail.

They were plotting these 2000 values on normal probability paper (QQ-plot), fitting a line, and extrapolating along the line far beyond the data.

For example, they wanted to know the chance of exceeding $\mu + 5\sigma$
($\approx 3 \times 10^{-7}$ for a normal distribution)
Certainly none of the 2000 observed cases would go there.

I questioned normality, asked for an example/sample of 2000 data points.
To my surprise they passed the normality test, which made me suspicious.

Their simulations had an internal switch, that produced normal data
and bypassed the costly simulations.

Appendix A

The following five slides prove the distribution result concerning the sum of independent normal random variables.

The proof is purely geometric.

$\alpha Z_1 + \beta Z_2 \sim \mathcal{N}(0,1)$ for $\alpha^2 + \beta^2 = 1$

The normal convolution result on the previous slide follows by induction from the following special case, which allows a simple and elegant proof.

Z_1 and Z_2 i.i.d. $\sim \mathcal{N}(0,1)$ and $\alpha^2 + \beta^2 = 1 \quad \Longrightarrow \quad \alpha Z_1 + \beta Z_2 \sim \mathcal{N}(0,1)$.

The crucial property that makes this proof possible is:

The joint density of (Z_1, Z_2) has circular symmetry around $(0,0)$

$$f(z_1, z_2) = \frac{1}{2\pi} \exp\left(- \frac{z_1^2 + z_2^2}{2} \right) \,,$$

i.e., points with same distance from $(0,0)$ have the same density.

Geometric Meaning of $\alpha Z_1 + \beta Z_2$ with $\alpha^2 + \beta^2 = 1$

$$\mathbf{Z} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = Z_1 \begin{pmatrix} 1 \\ 0 \end{pmatrix} + Z_2 \begin{pmatrix} 0 \\ 1 \end{pmatrix} = Z_1\, \mathbf{e}_1 + Z_2\, \mathbf{e}_2$$

$\mathbf{e}'_1 = (1,0)$ and $\mathbf{e}'_2 = (0,1)$ are the canonical orthonormal basis vectors in R^2.

$\mathbf{f}'_1 = (\alpha,\beta)$ and $\mathbf{f}'_2 = c(-\beta,\alpha)$ are also orthonormal basis vectors in R^2.

$\mathbf{f}'_1\mathbf{f}_1 = \alpha^2+\beta^2 = 1, \quad \mathbf{f}'_2\mathbf{f}_2 = (-\beta)^2+\alpha^2 = 1, \quad \mathbf{f}'_1\mathbf{f}_2 = \alpha(-\beta)+\beta\alpha = 0$

$$\mathbf{Z} = V_1\,\mathbf{f}_1 + V_2\,\mathbf{f}_2 \qquad \Longrightarrow \qquad \mathbf{f}'_1\mathbf{Z} = \alpha Z_1 + \beta Z_2 = \mathbf{f}'_1(V_1\,\mathbf{f}_1 + V_2\,\mathbf{f}_2) = V_1$$

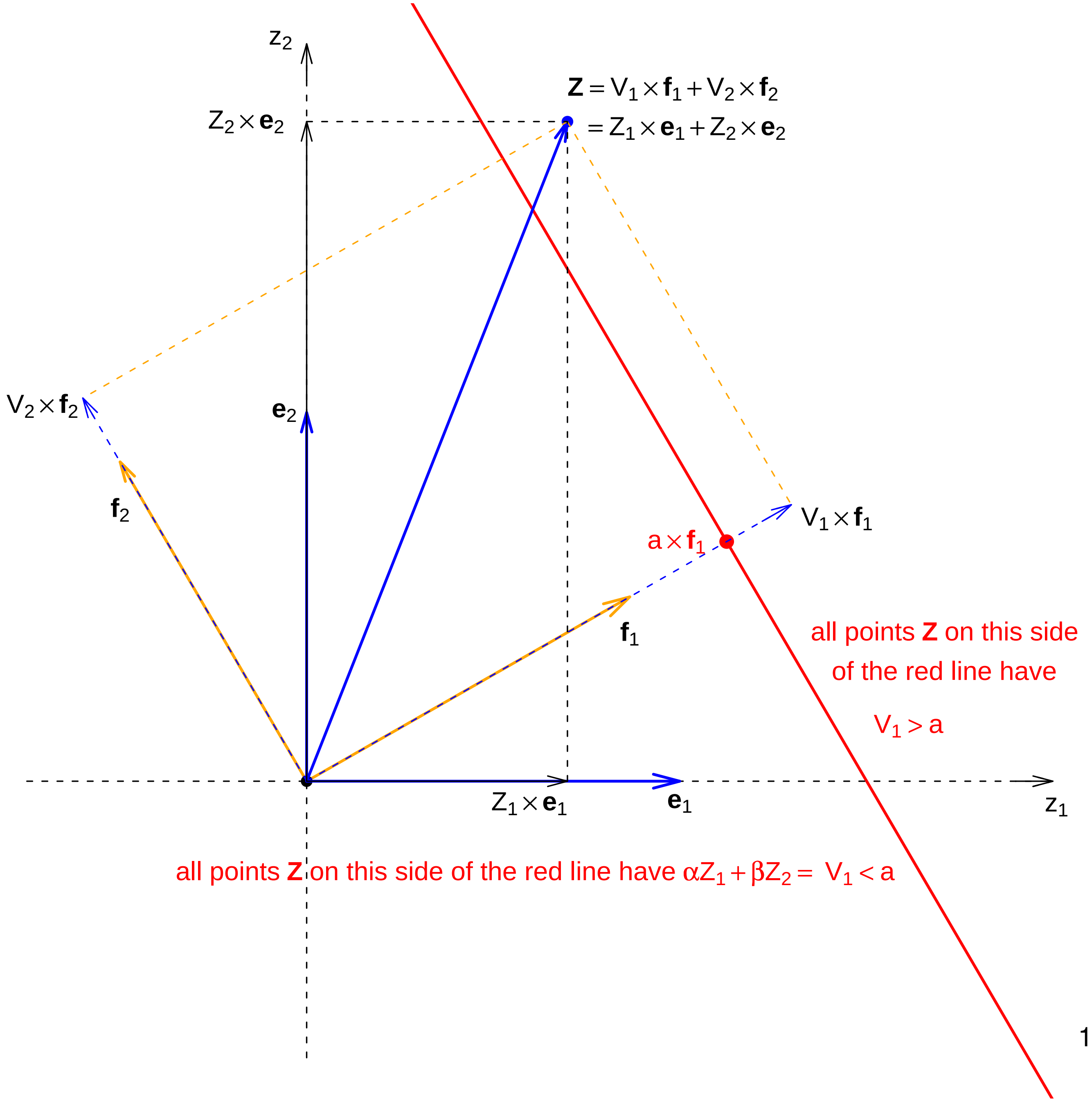
and $\qquad \mathbf{f}'_2\mathbf{Z} = -\beta Z_1 + \alpha Z_2 = \mathbf{f}'_2(V_1\,\mathbf{f}_1 + V_2\,\mathbf{f}_2) = V_2$

Thus $V_1 = \alpha Z_1 + \beta Z_2$ is the projection of $\mathbf{Z}$ onto the $\mathbf{f}_1$ direction

Correspondingly, $V_2 = -\beta Z_1 + \alpha Z_2$ is the projection of $\mathbf{Z}$ onto the $\mathbf{f}_2$ direction.

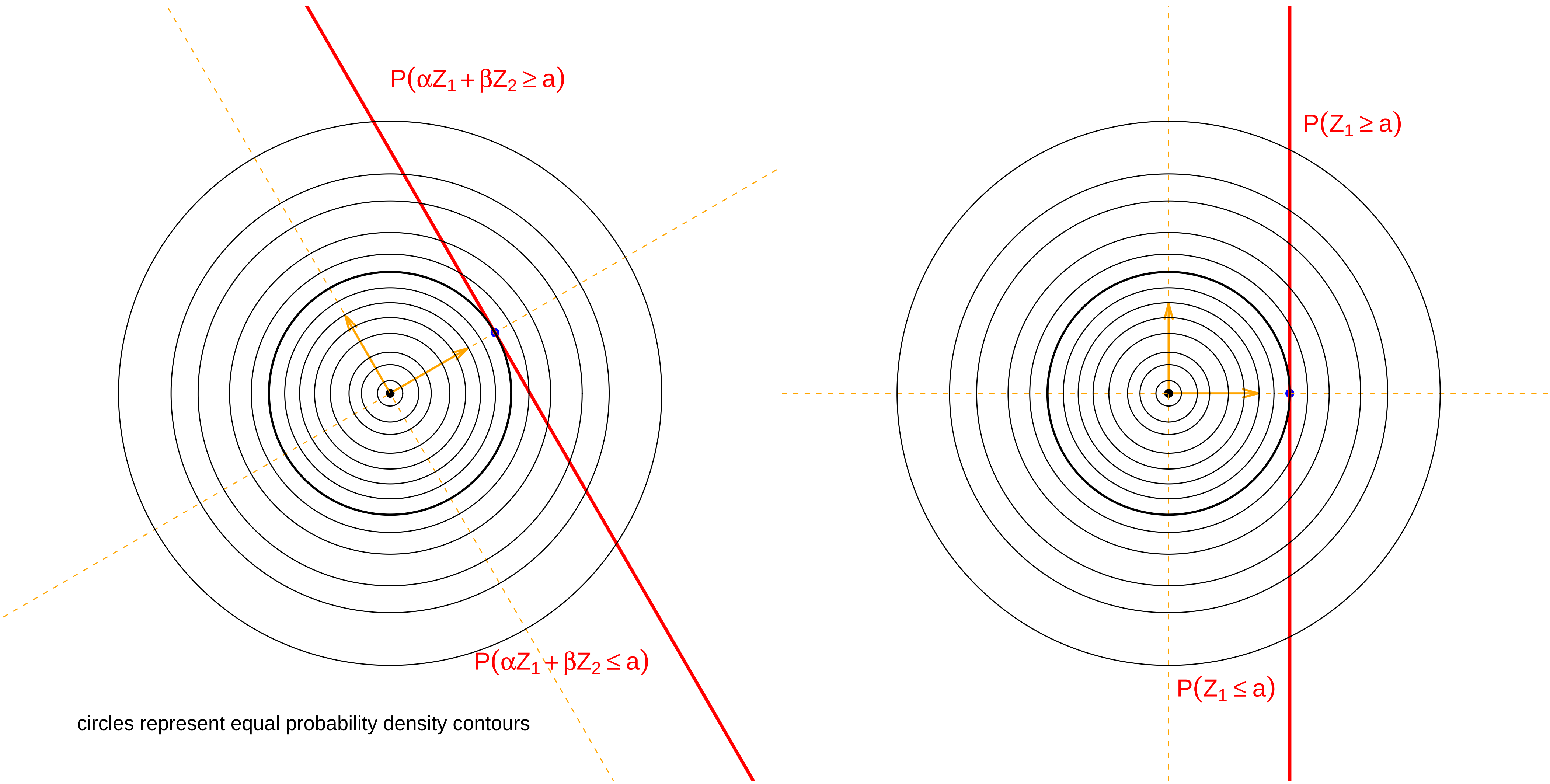
$$V_1\,\mathbf{f}_1 + V_2\,\mathbf{f}_2 = (\alpha Z_1 + \beta Z_2)\begin{pmatrix} \alpha \\ \beta \end{pmatrix} + (-\beta Z_1 + \alpha Z_2)\begin{pmatrix} -\beta \\ \alpha \end{pmatrix} = \begin{pmatrix} Z_1 \\ Z_2 \end{pmatrix} = \mathbf{Z}$$

Two Basis Representations of $\mathbf{Z}$



z_2
$Z_2 \times \mathbf{e}_2$
$\mathbf{Z} = V_1 \times \mathbf{f}_1 + V_2 \times \mathbf{f}_2$
$= Z_1 \times \mathbf{e}_1 + Z_2 \times \mathbf{e}_2$
$V_2 \times \mathbf{f}_2$
$\mathbf{e}_2$
$\mathbf{f}_2$
$V_1 \times \mathbf{f}_1$
$a \times \mathbf{f}_1$
$\mathbf{f}_1$
all points **Z** on this side
of the red line have
$V_1 > a$
$Z_1 \times \mathbf{e}_1$
$\mathbf{e}_1$
z_1
all points **Z** on this side of the red line have $\alpha Z_1 + \beta Z_2 = V_1 < a$

$$P(\alpha Z_1 + \beta Z_2 \leq a) = P(Z_1 \leq a)$$



circles represent equal probability density contours

$$\Longrightarrow X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)$$

The final step is

$$\frac{X_1 + X_2 - (\mu_1 + \mu_2)}{\sqrt{\sigma_1^2 + \sigma_2^2}} = \frac{\sigma_1}{\sqrt{\sigma_1^2 + \sigma_2^2}} \frac{X_1 - \mu_1}{\sigma_1} + \frac{\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}} \frac{X_2 - \mu_2}{\sigma_2} = \alpha Z_1 + \beta Z_2 \sim \mathcal{N}(0,1)$$

since $\alpha = \dfrac{\sigma_1}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ and $\beta = \dfrac{\sigma_2}{\sqrt{\sigma_1^2 + \sigma_2^2}}$ satisfy $\alpha^2 + \beta^2 = 1$

and $Z_1 = \dfrac{X_1 - \mu_1}{\sigma_1}$, $Z_2 = \dfrac{X_2 - \mu_2}{\sigma_2}$ are independent and $\sim \mathcal{N}(0,1)$

$$\Longrightarrow \qquad X_1 + X_2 \sim \mathcal{N}(\mu_1 + \mu_2, \sigma_1^2 + \sigma_2^2)\ .$$

and by induction $X_1 + \ldots + X_n \sim \mathcal{N}(\mu_1 + \ldots + \mu_n, \sigma_1^2 + \ldots + \sigma_n^2)$

Appendix B

The following six slides prove the distributional properties of $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ when $X_1, \ldots, X_n$ is a random sample from $\mathcal{N}(\mu, \sigma^2)$, namely

- $\bar{X}$ and $\sum_{i=1}^{n}(X_i - \bar{X})^2$ are statistically independent
- $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$
- $\sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$.

Rotational Symmetry of $(Z_1, \ldots, Z_n)$-Distribution

Assume that $(Z_1, \ldots, Z_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Then $(Z_1, \ldots, Z_n)$ has joint density

$$h(\mathbf{z}) = h(z_1, \ldots, z_n) = \varphi(z_1) \times \ldots \times \varphi(z_n) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \times \exp\left(-\frac{1}{2}\sum_{i=1}^{n} z_i^2\right)$$

Points equidistant from the origin (i.e., with constant $\sum z_i^2$) have same density.

Note

$$\mathbf{z} = \begin{pmatrix} z_1 \\ \vdots \\ z_n \end{pmatrix} = z_1 \begin{pmatrix} 1 \\ \vdots \\ 0 \end{pmatrix} + \ldots + z_n \begin{pmatrix} 0 \\ \vdots \\ 1 \end{pmatrix} = z_1 \mathbf{e}_1 + \ldots + z_n \mathbf{e}_n$$

$z_1, \ldots, z_n$ are the coordinates/coefficients with respect to the basis vectors $\mathbf{e}_1, \ldots, \mathbf{e}_n$.

Orthonormal Transformation Preserves i.i.d.$\mathcal{N}(0,1)$

Suppose we have another (rotated) orthonormal basis $\mathbf{f}_1,\ldots,\mathbf{f}_n$,

with $\mathbf{f}_i{}'\mathbf{f}_i = 1$ and $\mathbf{f}_i'\mathbf{f}_j = 0$ for $i \neq j$.

We reexpress $\mathbf{z}$ in terms of this basis, i.e., $\mathbf{z} = v_1\mathbf{f}_1 + \ldots + v_n\mathbf{f}_n$ with $\mathbf{z}'\mathbf{f}_i = v_i$.

$v_1,\ldots,v_n$ are the coordinates of the same vector $\mathbf{z}$ with respect to $\mathbf{f}_1,\ldots,\mathbf{f}_n$.

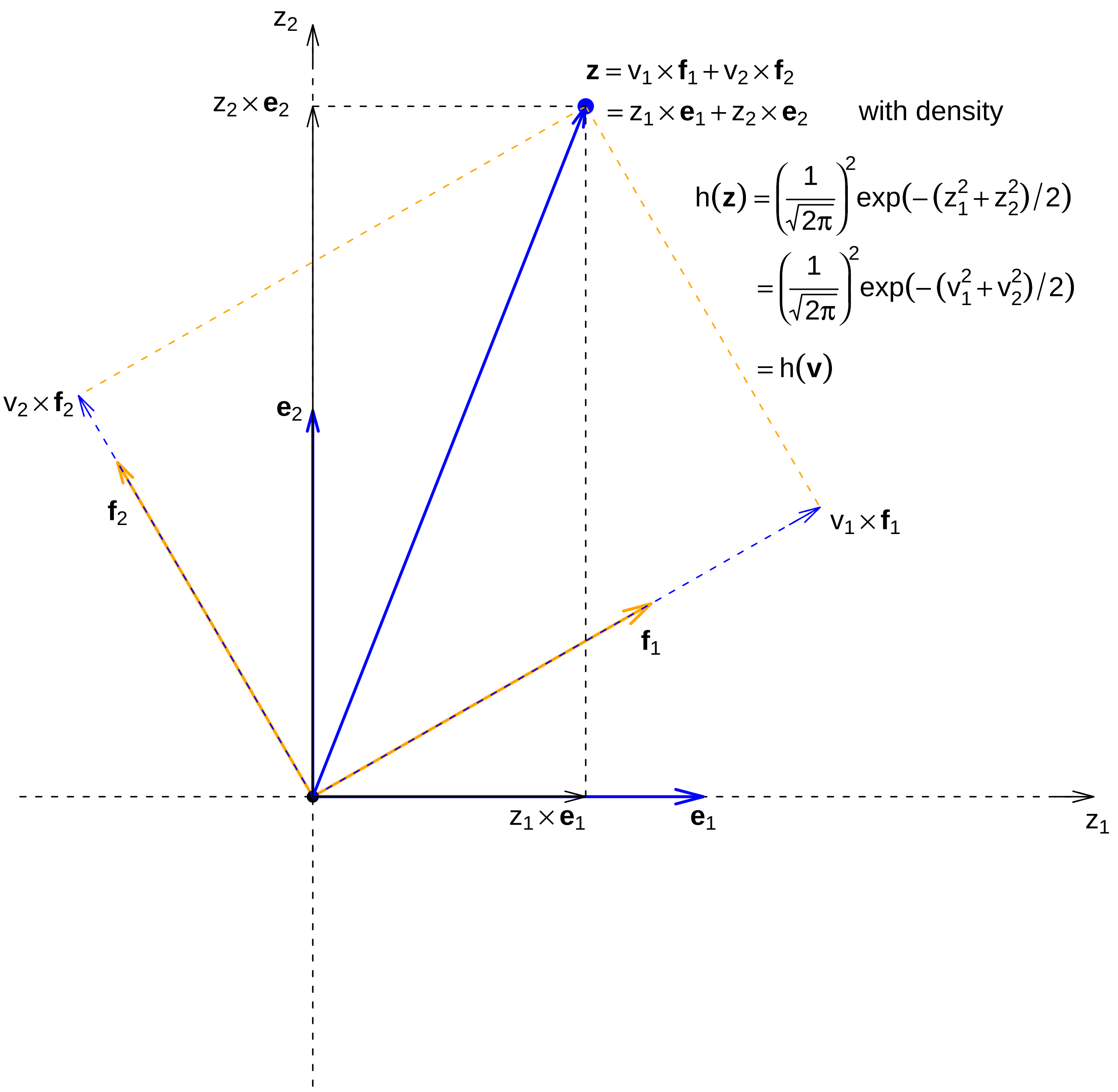
$$\Longrightarrow \quad \sum z_i^2 = (z_1\mathbf{e}_1 + \ldots + z_n\mathbf{e}_n)'(z_1\mathbf{e}_1 + \ldots + z_n\mathbf{e}_n)$$
$$= \mathbf{z}'\mathbf{z} = (v_1\mathbf{f}_1 + \ldots + v_n\mathbf{f}_n)'(v_1\mathbf{f}_1 + \ldots + v_n\mathbf{f}_n) = \sum v_i^2$$

Associate the density $h(\mathbf{z})$ with $\mathbf{v} = \mathbf{v}(\mathbf{z})$

$$h(\mathbf{z}) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_i z_i^2\right) = \left(\frac{1}{\sqrt{2\pi}}\right)^n \exp\left(-\frac{1}{2}\sum_i v_i^2\right) = h(\mathbf{v}(\mathbf{z})) = h(\mathbf{v})$$

thus $V_1,\ldots,V_n \overset{\text{i.i.d.}}{\sim} \mathcal{N}(0,1)$.

Two Basis Representations of $\mathbf{z}$



$$\mathbf{z} = v_1 \times \mathbf{f}_1 + v_2 \times \mathbf{f}_2$$
$$= z_1 \times \mathbf{e}_1 + z_2 \times \mathbf{e}_2 \quad \text{with density}$$

$$h(\mathbf{z}) = \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp(-(z_1^2 + z_2^2)/2)$$
$$= \left(\frac{1}{\sqrt{2\pi}}\right)^2 \exp(-(v_1^2 + v_2^2)/2)$$
$$= h(\mathbf{v})$$

z_2
$z_2 \times \mathbf{e}_2$
$\mathbf{e}_2$
$v_2 \times \mathbf{f}_2$
$\mathbf{f}_2$
$v_1 \times \mathbf{f}_1$
$\mathbf{f}_1$
$z_1 \times \mathbf{e}_1$
$\mathbf{e}_1$
z_1

Distribution of $\bar{Z}$ and $\sum(Z_i - \bar{Z})^2$

Suppose we choose $\mathbf{f}_1' = (1/\sqrt{n}, \ldots, 1/\sqrt{n})$ and choose orthonormal vectors for the other $\mathbf{f}_i$. $\Longleftarrow$ Gram-Schmidt orthogonalization based on the basis $\mathbf{f}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$.

Then $\mathbf{z}'\mathbf{f}_1 = (v_1\mathbf{f}_1 + \ldots + v_n\mathbf{f}_n)'\mathbf{f}_1 = v_1 = \sum z_i/\sqrt{n} = \sqrt{n}\,\bar{z}$.

$$\sum_{i=1}^{n} z_i^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2 + n\bar{z}^2 = \sum_{i=1}^{n} (z_i - \bar{z})^2 + v_1^2 = \sum_{i=1}^{n} v_i^2 \Longrightarrow \sum_{i=1}^{n} (z_i - \bar{z})^2 = \sum_{i=2}^{n} v_i^2 \,.$$

$$\Longrightarrow \qquad \sqrt{n}\bar{Z} = V_1 \sim \mathcal{N}(0,1) \qquad \text{or} \qquad \bar{Z} \sim \mathcal{N}(0, 1/n)$$

is independent of

$$\sum_{i=1}^{n} (Z_i - \bar{Z})^2 = \sum_{i=2}^{n} V_i^2 \sim \chi^2_{n-1} \,.$$

Review of Gram-Schmidt orthogonalization

$\mathbf{f}_1, \mathbf{e}_2, \ldots, \mathbf{e}_n$ are a basis of R^n, since $\mathbf{f}_1 = (1, \ldots, 1)'/\sqrt{n} \neq \sum_{i=2}^n a_i \mathbf{e}_i$ for any $(a_2, \ldots, a_n)$.

We get orthogonal basis vectors $\mathbf{f}_i$ successively as follows: $\mathbf{f}_1 = \mathbf{f}_1$ and

$$\mathbf{f}_2 = \mathbf{e}_1 - a_{21}\mathbf{f}_1 \quad \Longrightarrow \quad \mathbf{f}_1'\mathbf{f}_2 = \mathbf{f}_1'\mathbf{e}_1 - a_{21} = 0 \quad \Longrightarrow \quad a_{21} = \mathbf{f}_1'\mathbf{e}_1$$

$$\mathbf{f}_3 = \mathbf{e}_2 - a_{31}\mathbf{f}_1 - a_{32}\mathbf{f}_2 \quad \Longrightarrow \quad \mathbf{f}_1'\mathbf{f}_3 = \mathbf{f}_1'\mathbf{e}_2 - a_{31} = 0 \text{ and } \mathbf{f}_2'\mathbf{f}_3 = \mathbf{f}_2'\mathbf{e}_2 - a_{32} = 0$$

from previously constructed orthogonality $\mathbf{f}_1'\mathbf{f}_2 = \mathbf{f}_2'\mathbf{f}_1 = 0$, thus $a_{3i} = \mathbf{f}_i'\mathbf{e}_2,\ i = 1, 2.$

Next $\mathbf{f}_4 = \mathbf{e}_3 - a_{41}\mathbf{f}_1 - a_{42}\mathbf{f}_2 - a_{43}\mathbf{f}_3$ and multiplying this equation respectively by $\mathbf{f}_1', \mathbf{f}_2', \mathbf{f}_3'$ and setting to zero we get

$$a_{41} = \mathbf{f}_1'\mathbf{e}_3\,, \quad a_{42} = \mathbf{f}_2'\mathbf{e}_3\,, \quad \text{and} \quad a_{43} = \mathbf{f}_3'\mathbf{e}_3 \qquad \text{and so on.}$$

$\mathbf{f}_i \Big/ \sqrt{\mathbf{f}_i'\mathbf{f}_i}\,,\ \ i = 1, \ldots, n$ are then our orthonormal basis vectors.

Distribution of $\bar{X}$ and $\sum(X_i - \bar{X})^2$

Assume that $(X_1, \ldots, X_n) \overset{\text{i.i.d.}}{\sim} \mathcal{N}(\mu, \sigma^2)$.

Then $(Z_1, \ldots, Z_n)$ with $Z_i = (X_i - \mu)/\sigma$ are $\overset{\text{i.i.d.}}{\sim} \mathcal{N}(0, 1)$.

From the previous result we have that $\sqrt{n}\bar{Z} = \sqrt{n}(\bar{X} - \mu)/\sigma \sim \mathcal{N}(0, 1)$ and thus $\bar{X} = \mu + \sigma\bar{Z} \sim \mathcal{N}(\mu, \sigma^2/n)$ and it is independent of

$$\sum_{i=1}^{n}(Z_i - \bar{Z})^2 = \sum_{i=1}^{n}((X_i - \mu)/\sigma - (\bar{X} - \mu)/\sigma)^2 = \sum_{i=1}^{n}(X_i - \bar{X})^2/\sigma^2 \sim \chi^2_{n-1}$$

or

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \text{ has the same distribution as } \sigma^2 C_{n-1} \text{ where } C_{n-1} \sim \chi^2_{n-1} \,.$$

$$\sum_{i=1}^{n}(X_i - \bar{X})^2 \quad \text{and} \quad \bar{X} \quad \text{are independent .}$$