Sampling Distributions of Statistics

Corresponds to Chapter 5 of Tamhane and Dunlop

Slides prepared by Elizabeth Newton (MIT),
with some slides by Jacqueline Telford
(Johns Hopkins University)

Sampling Distributions

Definitions and Key Concepts

• A sample statistic used to estimate an unknown population parameter is called an estimate.

• The discrepancy between the estimate and the true parameter value is known as sampling error.

• A statistic is a random variable with a probability distribution, called the sampling distribution, which is generated by repeated sampling.

• We use the sampling distribution of a statistic to assess the sampling error in an estimate.

Random Sample

- Definition 5.11, page 201, Casella and Berger.
- How is this different from a simple random sample?
- For mutual independence, population must be very large or must sample with replacement.

Sample Mean and Variance

Sample Mean

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$$

Sample Variance

$$S^2 = \frac{\sum_{i=1}^{n} (X_i - \overline{X})^2}{n - 1}$$

How do the sample mean and variance vary in repeated samples of size n drawn from the population?

In general, difficult to find exact sampling distribution. However, see example of deriving distribution when all possible samples can be enumerated (rolling 2 dice) in sections 5.1 and 5.2. Note errors on page 168.

Properties of a sample mean and variance

See Theorem 5.2.2, page 268, Casella & Berger.

Distribution of Sample Means

- If the i.i.d. r.v.'s are
  - Bernoulli
  - Normal
  - Exponential

  The distributions of the sample means can be derived

  Sum of n i.i.d. Bernoulli(p) r.v.'s is Binomial(n,p)

  Sum of n i.i.d. Normal(μ,σ^2) r.v.'s is Normal($n\mu,n\sigma^2$)

  Sum of n i.i.d. Exponential(λ) r.v.'s is Gamma(λ,n)

Distribution of Sample Means

- Generally, the exact distribution is difficult to calculate.
- What can be said about the distribution of the sample mean when the sample is drawn from an arbitrary population?
- In many cases we can approximate the distribution of the sample mean when *n* is large by a normal distribution.
- **The famous Central Limit Theorem**

Central Limit Theorem

Let $X_1, X_2, \ldots, X_n$ be a random sample drawn from an **arbitrary** distribution with a finite mean μ and variance σ^2

As *n* goes to infinity, the sampling distribution of

$$\frac{\overline{X} - \mu}{\sigma / \sqrt{n}}$$ converges to the N(0,1) distribution.

Sometimes this theorem is given in terms of the sums:

$$\frac{\sum_{i=1}^{n} X_i - n\mu}{\sqrt{n}\,\sigma} \approx N(0,1)$$

Central Limit Theorem

Let $X_1 \ldots X_n$ be a random sample from an arbitrary distribution with finite mean μ and variance σ^2. As n increases

$$\frac{(\overline{X} - \mu)}{\sigma / \sqrt{n}} \approx N(0,1)$$

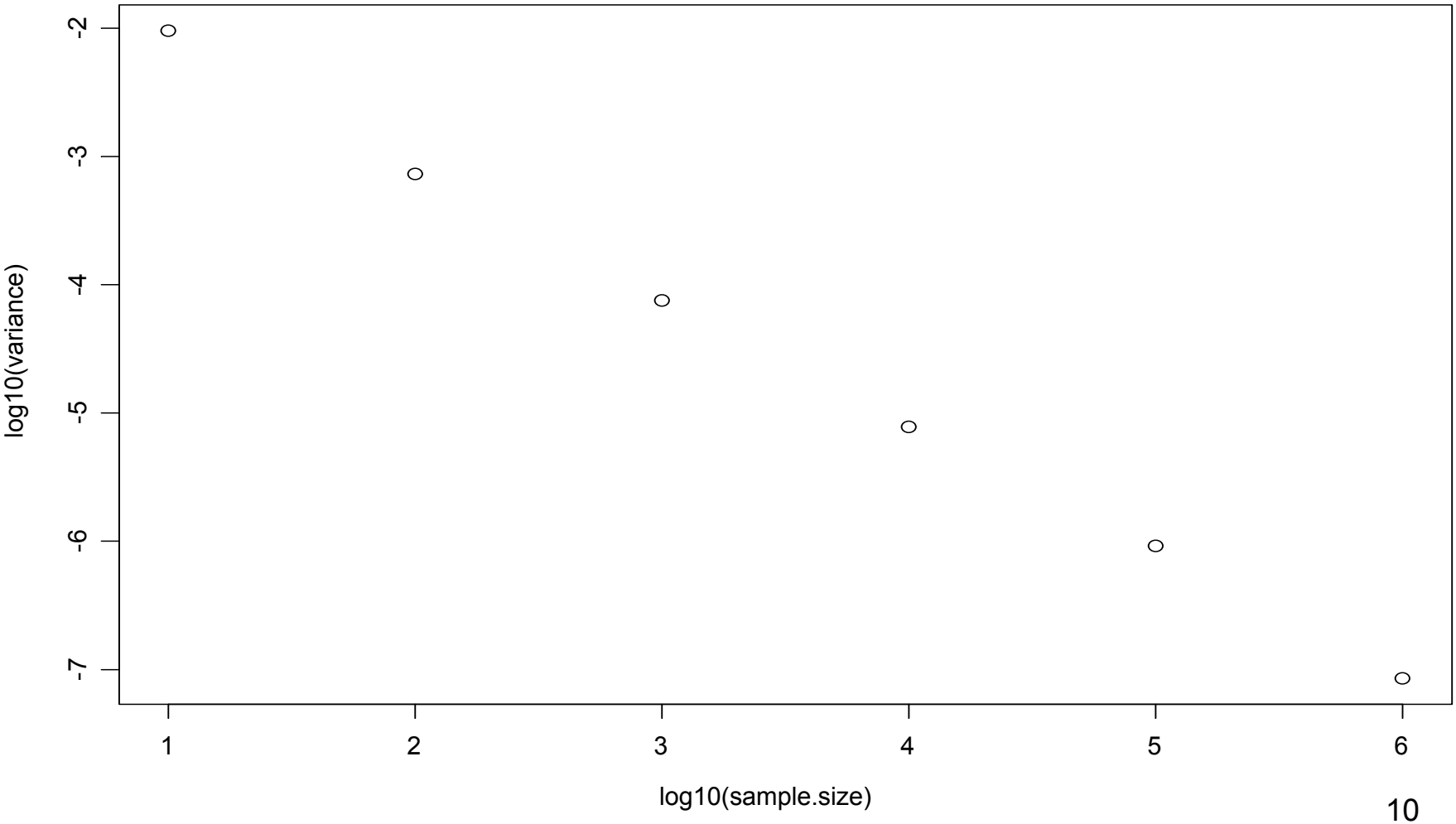
$$\Rightarrow \overline{X} \approx N(\mu, \frac{\sigma^2}{n})?$$

$$\Rightarrow \sum_{i=1}^{n} X_i \approx N(n\mu, n\sigma^2)?$$

What happens as n goes to infinity?

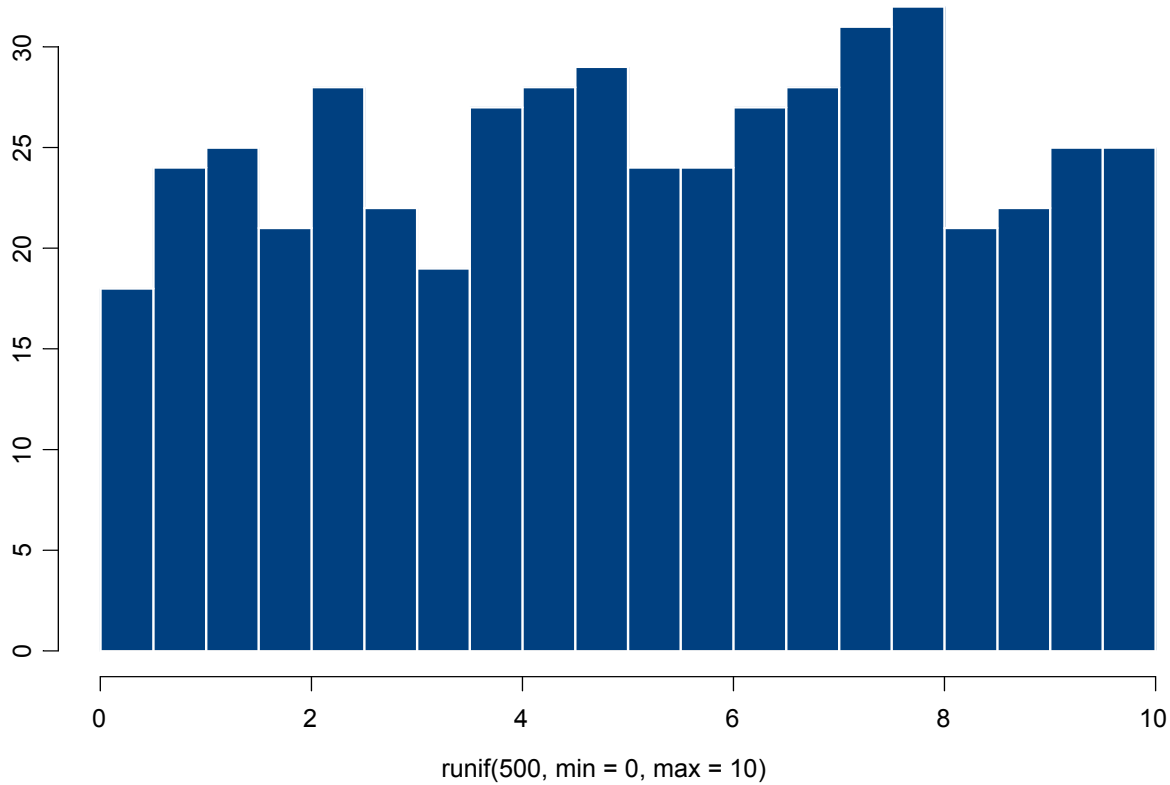
Variance of means from uniform distribution

sample size=10 to 10^6
number of samples=100

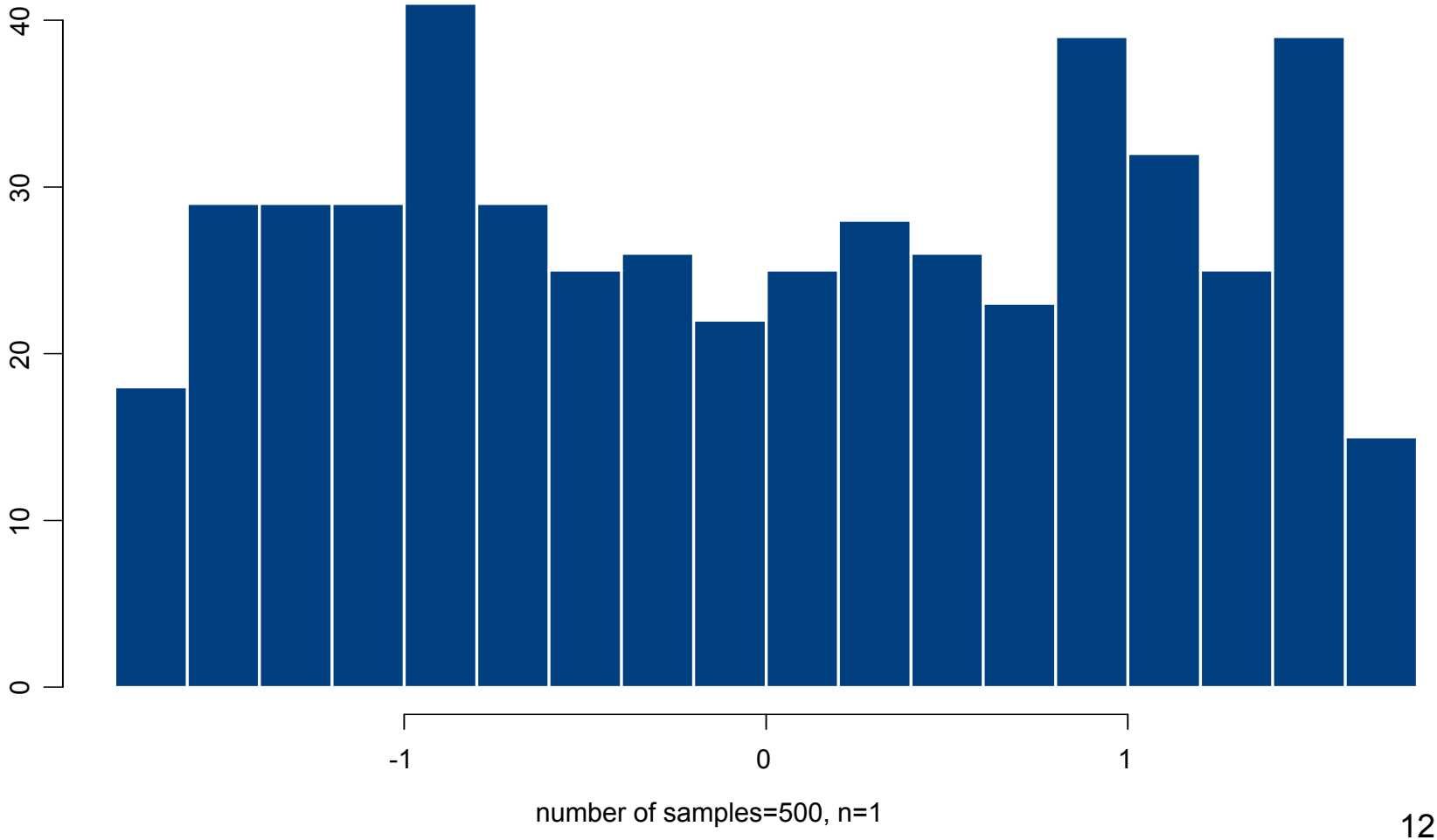


Example: Uniform Distribution

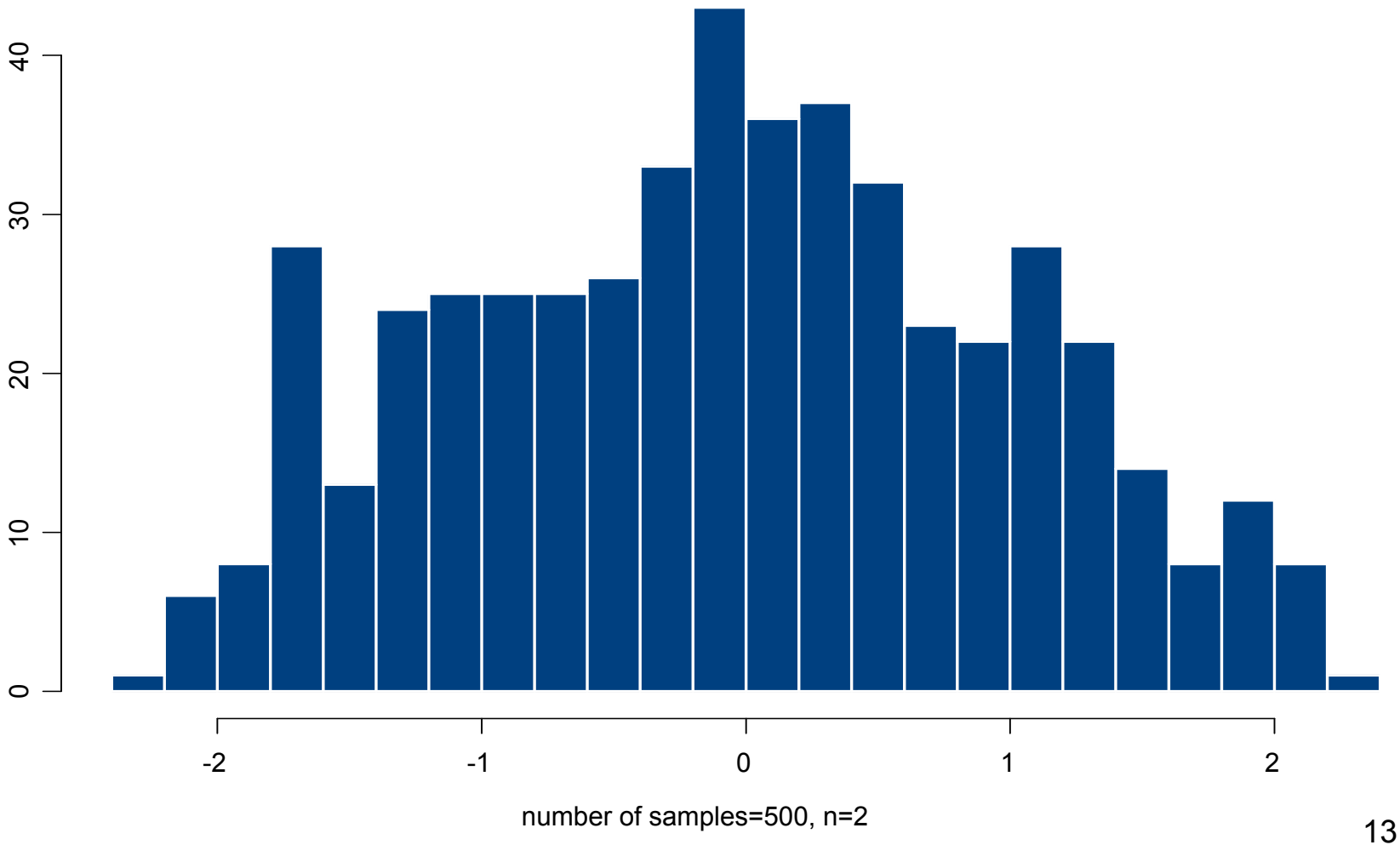
- $f(x \mid a, b) = 1 / (b-a)$, $a \le x \le b$
- $E\ X = (b+a)/2$
- $Var\ X = (b-a)^2/12$



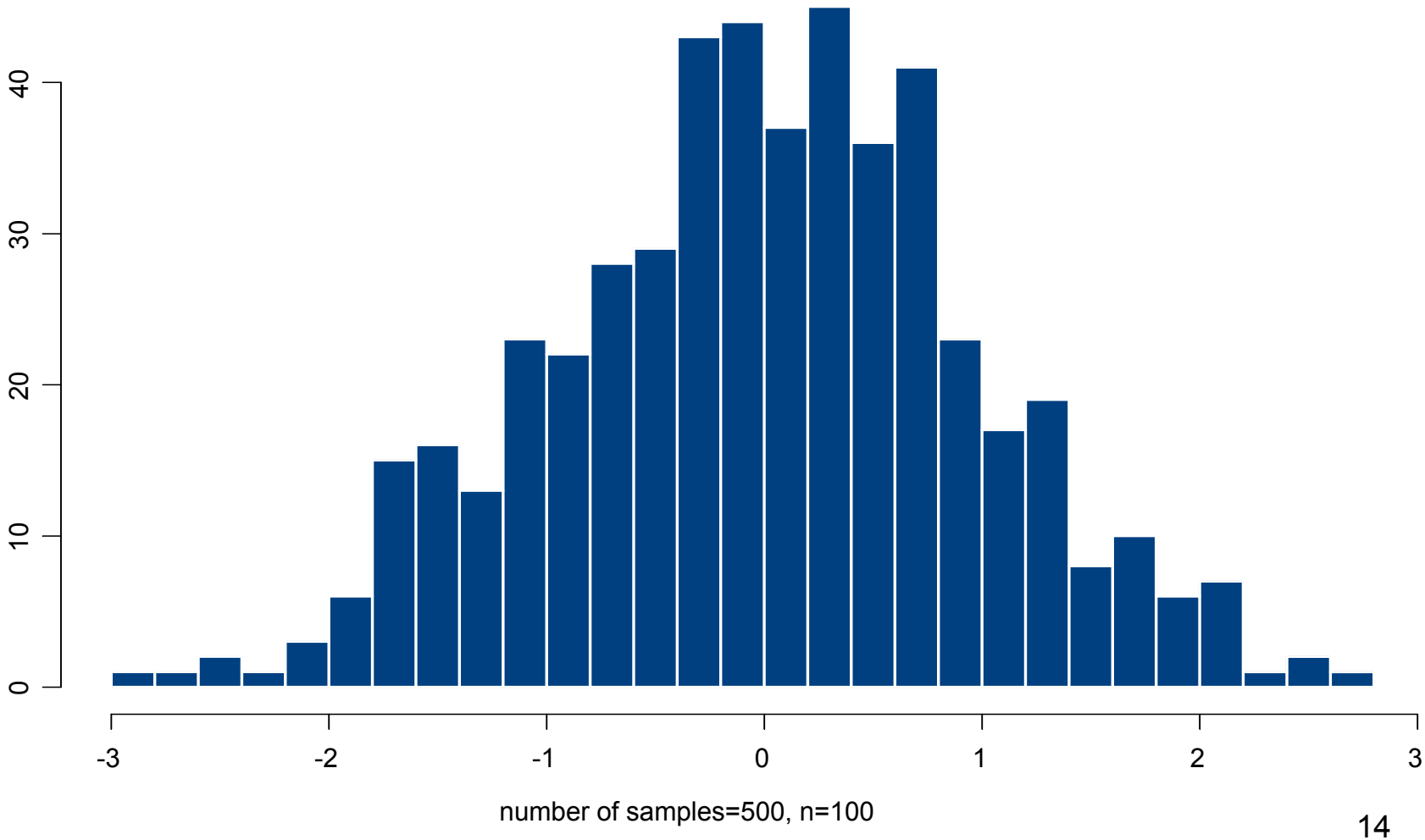
Standardized Means, Uniform Distribution
500 samples, n=1



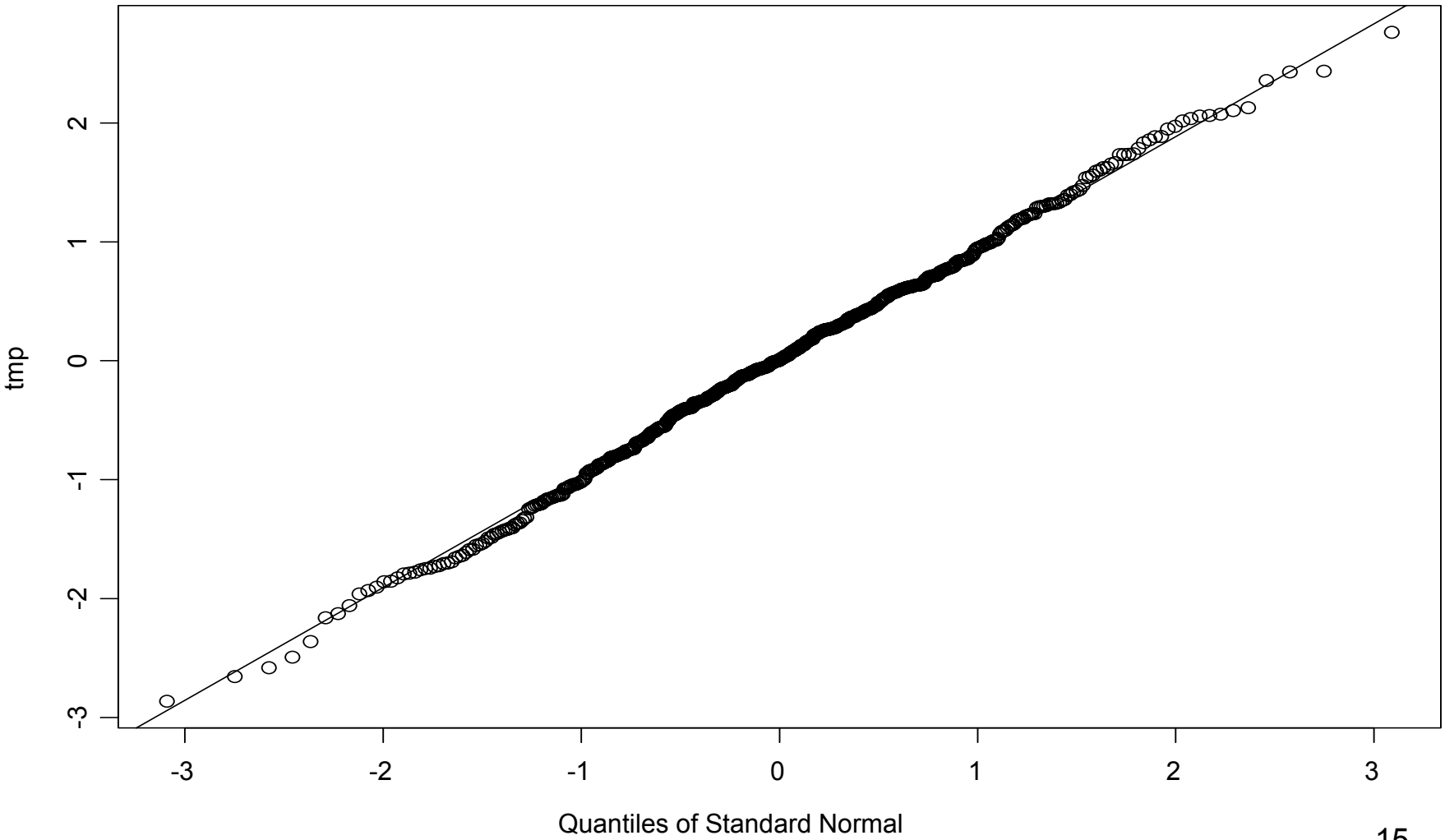
Standardized Means, Uniform Distribution 500 samples, n=2



Standardized Means, Uniform Distribution
500 samples, n=100



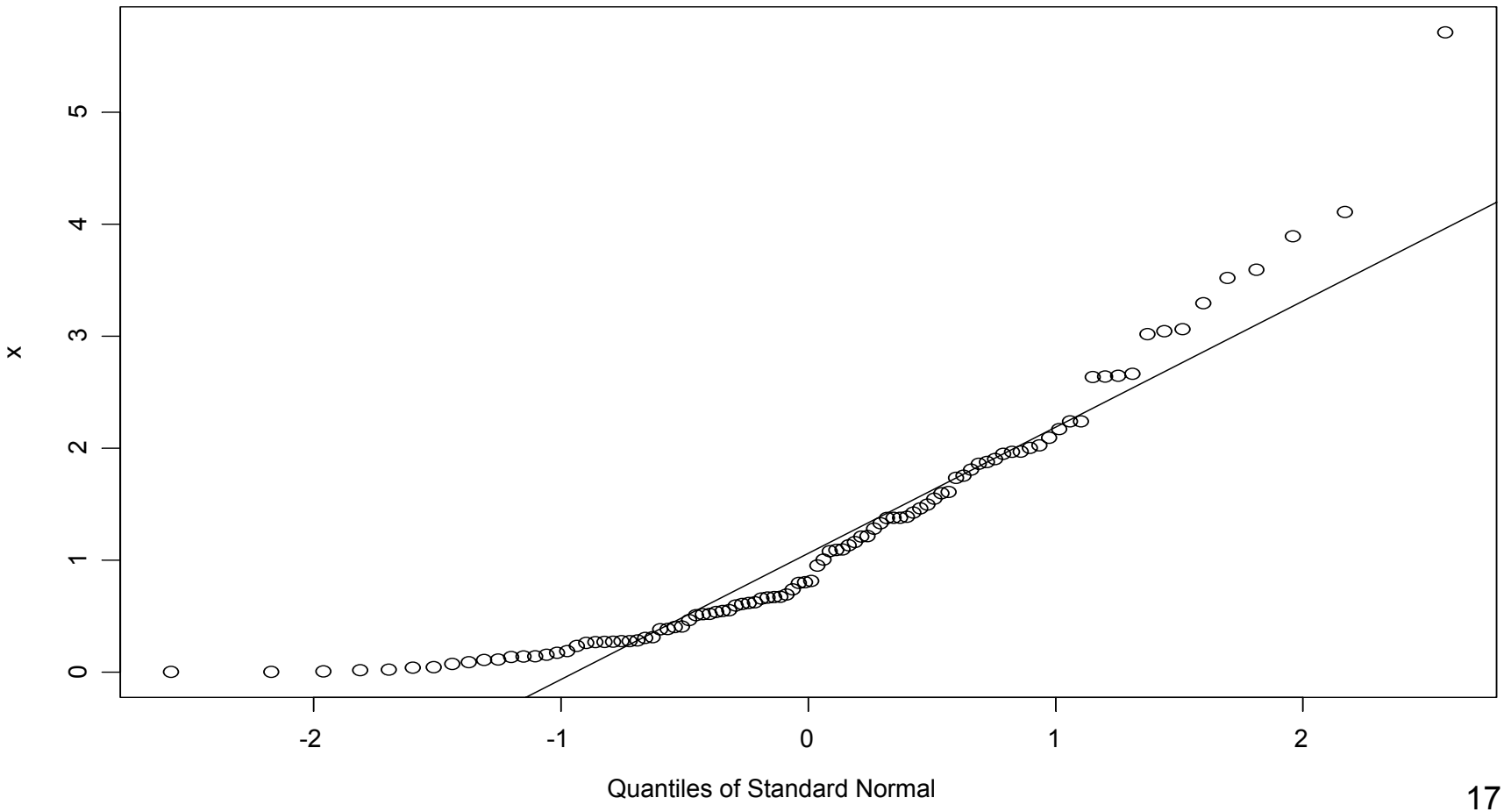
QQ (Normal) plot of means of 500 samples of size 100 from uniform distribution



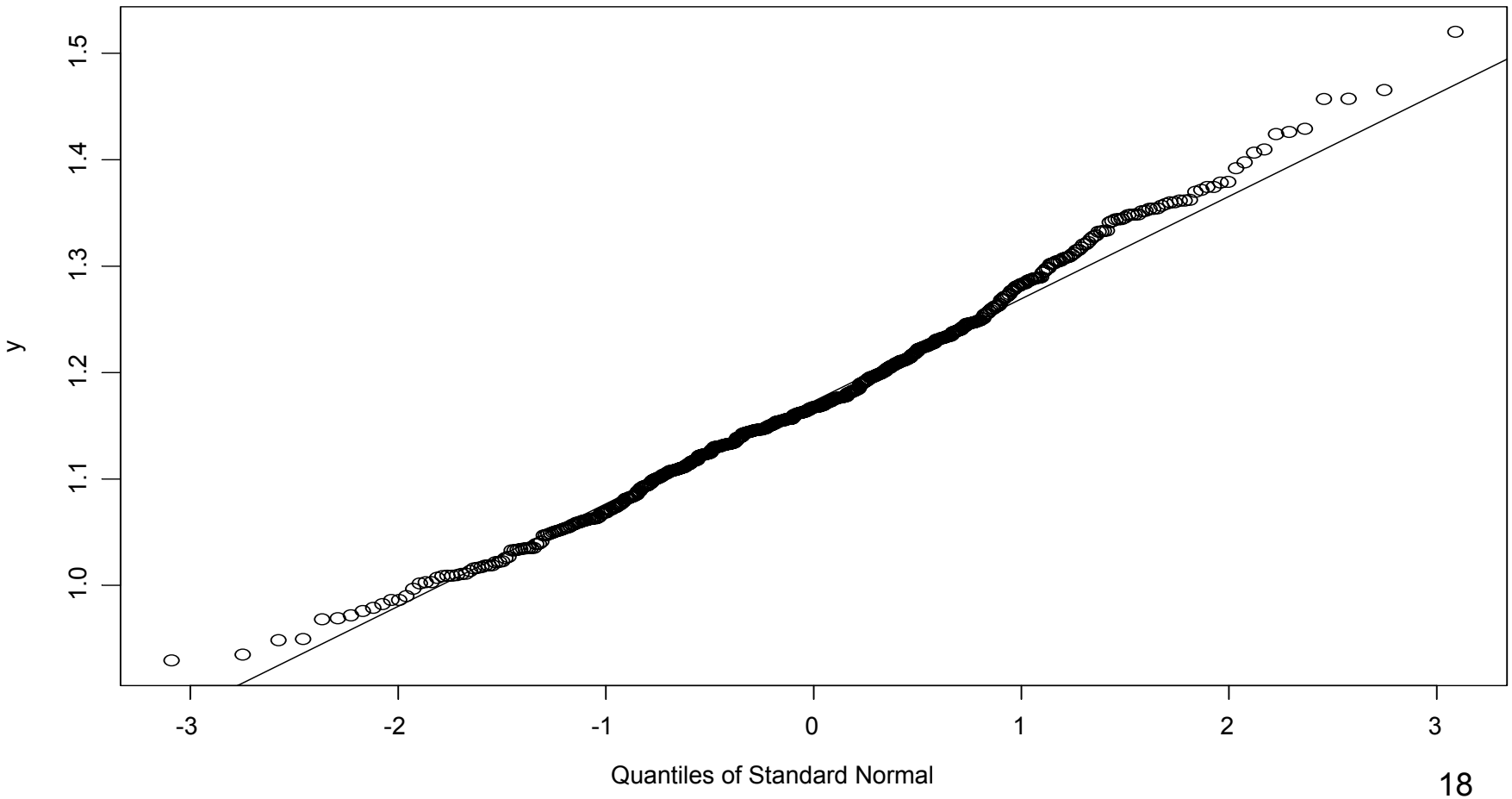
Bootstrap – sampling from the sample

- Previous slides have shown results for means of 500 samples (of size 100) from uniform distribution.
- Bootstrap takes just one sample of size 100 and then takes 500 samples (of size 100) with replacement from the sample.
- x<-runif(100)
- y<- mean(sample(x,100,replace=T))

Normal probability plot of sample of size 100 from exponential distribution



Normal probability plot of means of 500 bootstrap samples from sample of size 100 from exponential distribution



Law of Large Numbers and Central Limit Theorem

Both are asymptotic results about the sample mean:

• Law of Large Numbers (LLN) says that as n $\to \infty$, the sample mean converges to the population mean, i.e.,

$$\text{as n} \to \infty, \overline{X} - \mu \to 0$$

• Central Limit Theorem (CLT) says that as n $\to \infty$, also the distribution converges to Normal, i.e.,

$$\text{as n} \to \infty, \frac{\overline{X} - \mu}{\sigma / \sqrt{n}} \text{ converges to N(0,1)}$$

Normal Approximation to the Binomial

A binomial r.v. is the sum of i.i.d. Bernoulli r.v.'s so the CLT can be used to approximate its distribution.

Suppose that X is B(*n, p*). Then the mean of *X* is *np* and the variance of X is *np(1 - p)* .

By the CLT, we have: $\frac{X - np}{\sqrt{np(1-p)}} \approx N(0,1)$

$$\left[\begin{array}{l}\text{General} \\ \text{Formula}\end{array} = \frac{r.v. - E(r.v.)}{SD(r.v.)}\right]$$

How large a sample, *n*, do we need for the approximation to be good?

Rule of Thumb: $np \geq 10$ and $n(1-p) \geq 10$

For p=0.5, $np = n(1-p) = n(0.5) = 10 \Rightarrow n$ should be 20. (symmetrical)

For p=0.1 or 0.9, *np* or $n(1-p) = n(0.1) = 10 \Rightarrow n$ should be 100. (skewed)

- See Figures 5.2 and 5.3 and Example 5.3, pp.172-174

Continuity Correction

See Figure 5.4 for motivation.

$$P(X \le x) \cong \Phi\left(\frac{x + 0.5 - np}{\sqrt{np(1-p)}}\right)$$

$$P(X \ge x) \cong 1 - \Phi\left(\frac{x - 0.5 - np}{\sqrt{np(1-p)}}\right)$$

Exact Binomial Probability:

$P(X \le 8)$ = 0.2517

Normal approximation without Continuity Correction:

$P(X \le 8)$ = 0.1867

Normal approximation with Continuity Correction:

$P(X \le 8.5)$ = 0.2514 (much better agreement with exact calculation)

Sampling Distribution of the Sample Variance

$$S^2 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})^2}{n-1} \sim ?$$

There is no analog to the CLT for which gives an approximation for large samples for an arbitrary distribution.

The exact distribution for S^2 can be derived for X ~ i.i.d. Normal.

Chi-square distribution: For $\nu \geq 1$, let $Z_1, Z_2, \ldots, Z_\nu$ be i.i.d. N(0,1) and let $Y = Z_1^2 + Z_2^2 + \ldots + Z_\nu^2$.

The p.d.f. of Y can be shown to be $f(y) = \frac{1}{2^{\nu/2}\Gamma(\nu/2)} x^{(\nu/2)-1} e^{-y/2}$

This is known as the χ^2 distribution with ν degrees of freedom (d.f.) or Y ~ χ^2_ν.

- See Figures 5.5 and 5.6, pp. 176-177 and Table A.5, p.676

Distribution of the Sample Variance in the Normal Case

If $Z \sim N(0,1)$, then $Z^2 \sim \chi^2_1$

It can be shown that $\frac{(n-1)S^2}{\sigma^2} = \frac{S^2}{\sigma^2/(n-1)} \sim \chi^2_{n-1}$

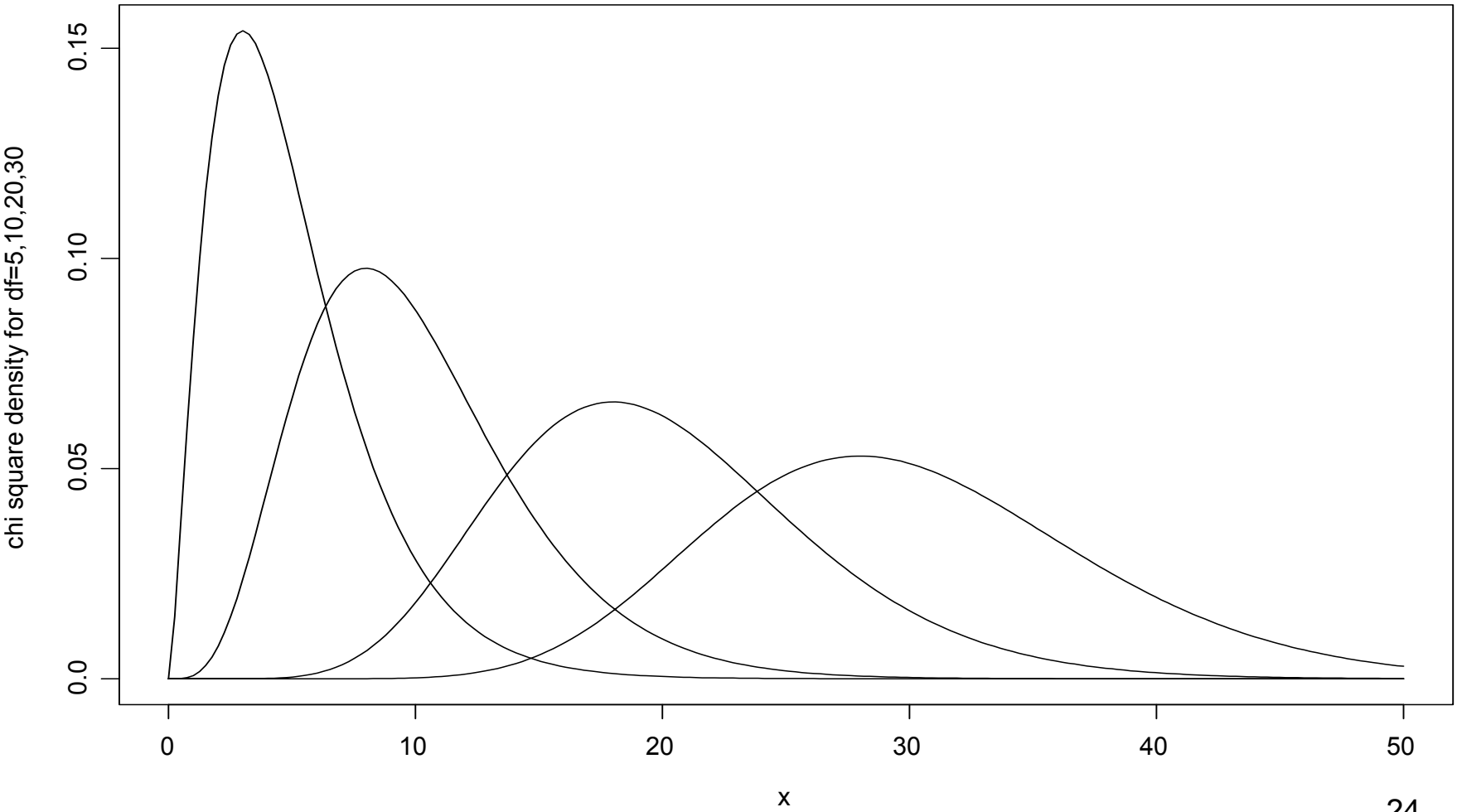
or equivalently $S^2 \sim \frac{\sigma^2 \chi^2_{n-1}}{n-1}$, a scaled χ^2

$E(S^2) = \sigma^2$ (is an unbiased estimator)

$Var(S^2) = \frac{2\sigma^4}{n-1}$

See Result 2 (p.179)

Chi-square distribution



Chi-Square Distribution Interesting Facts

- EX = v (degrees of freedom)
- Var X = 2v
- Special case of the gamma distribution with scale parameter=2, shape parameter=v/2.
- Chi-square variate with v d.f. is equal to the sum of the squares of v independent unit normal variates.

Student's *t*-Distribution

Consider a random sample $X_1, X_2, ..., X_n$ drawn from $N(\mu,\sigma^2)$.

It is known that $\dfrac{\bar{X}-\mu}{\sigma/\sqrt{n}}$ is exactly distributed as N(0,1).

$T = \dfrac{\bar{X}-\mu}{S/\sqrt{n}}$ is NOT distributed as N(0,1).

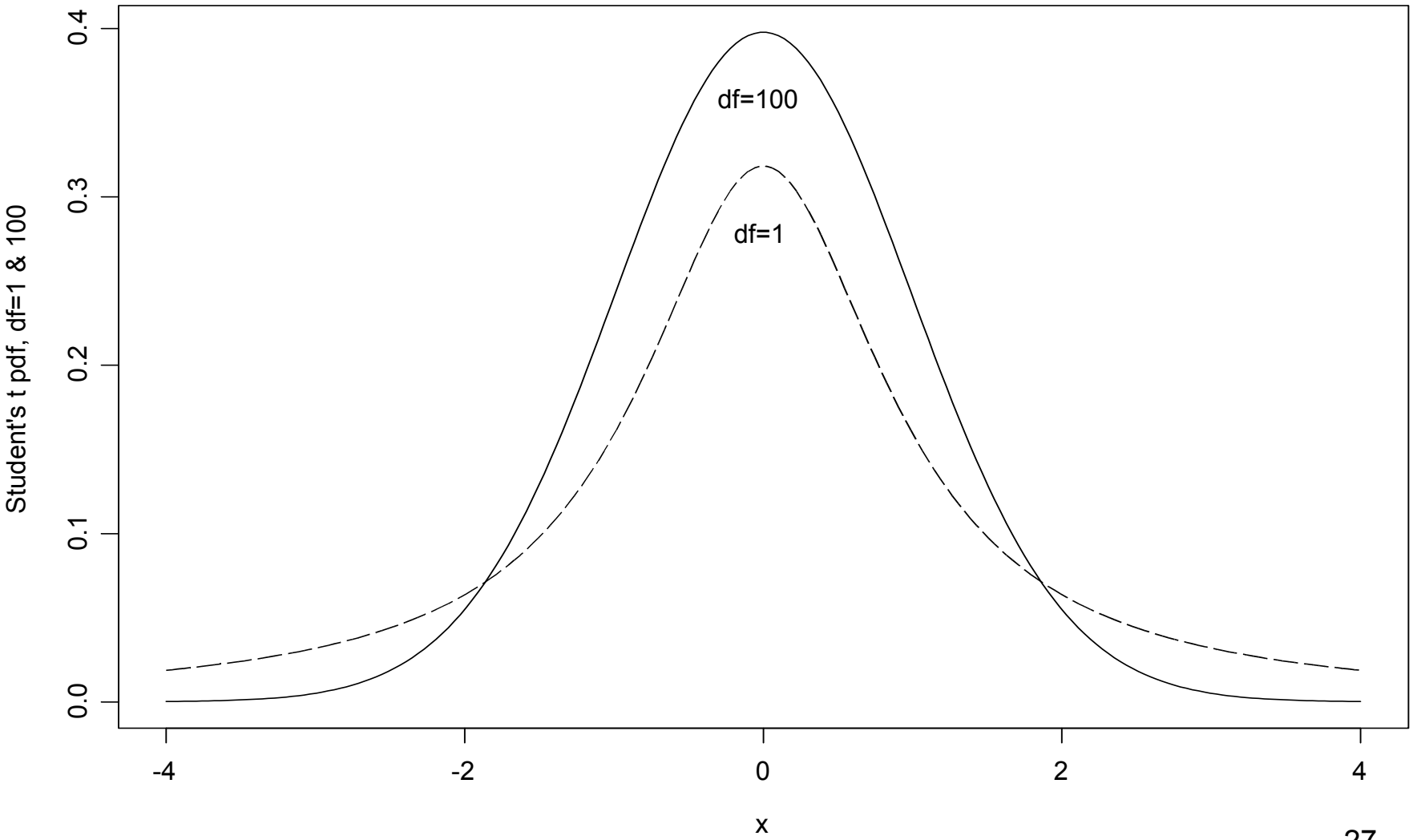
A different distribution for each ν = n-1 degrees of freedom (d.f.).

T is the ratio of a N(0,1) r.v. and sq.rt.(independent χ^2 divided by its d.f.)
- for derivation, see eqn 5.13, p.180, and its messy p.d.f., eqn 5.14

See Figure 5.7, Student's *t* p.d.f.'s for ν = 2, 10,and ∞, p.180

- See Table A.4, *t*-distribution table, p. 675
- See Example 5.6, milk cartons, p. 181

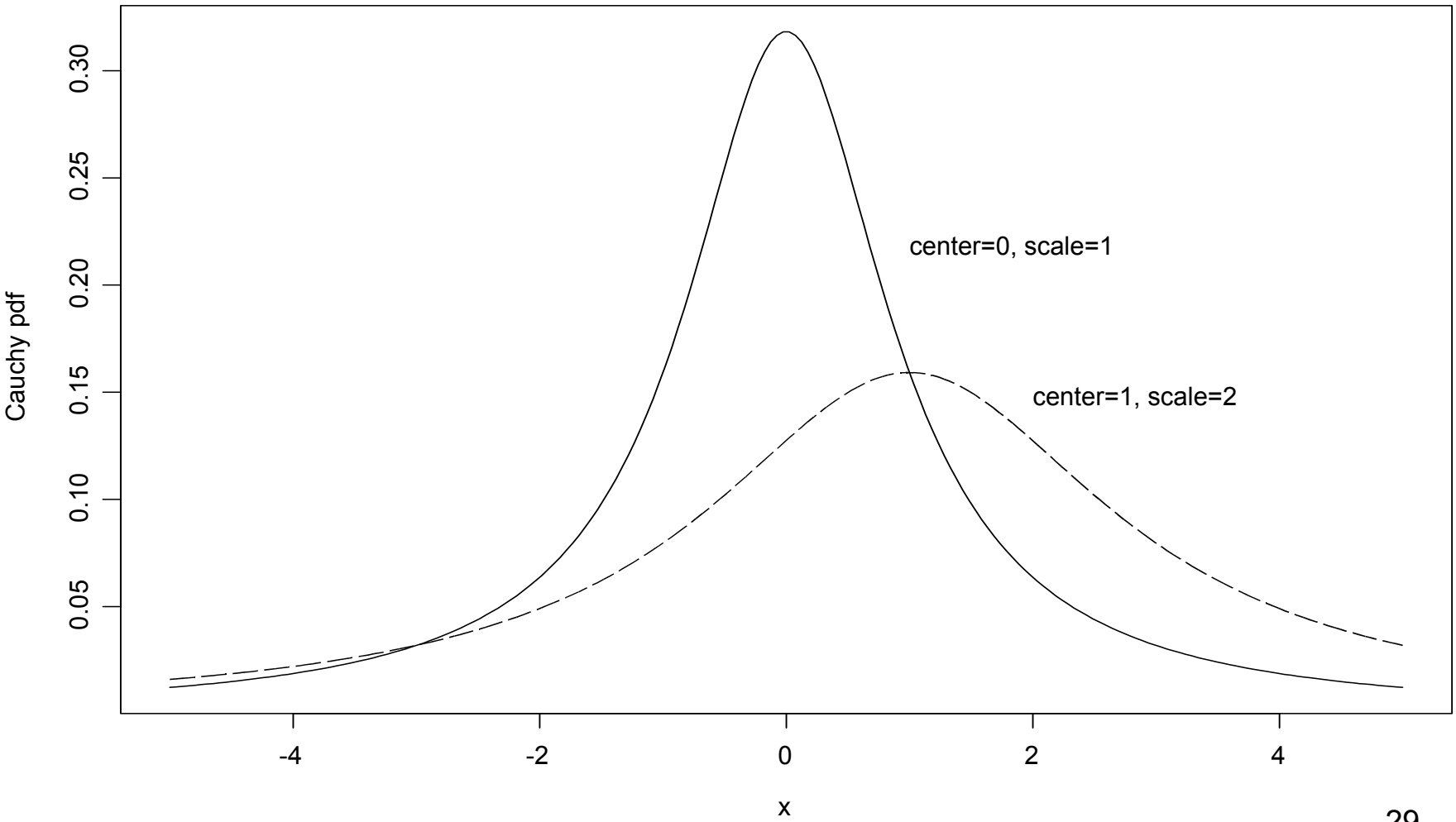
Student's t densities for df=1,100



Student's t Distribution Interesting Facts

- E X = 0, for v>1
- Var X = v/(v-2) for v>2
- Related to F distribution ($F_{1,v} = t^2_v$)
- As v tends to infinity t variate tends to unit normal
- If v=1 then t variate is standard Cauchy

Cauchy Distribution for center=0, scale=1 and center=1, scale=2



Cauchy Distribution Interesting Facts

$$f(x \mid a,b) = \{\pi b[1 + (\frac{x-a}{b})^2]\}^{-1}$$

- Parameters, a=center, b=scale
- Mean and Variance do not exist (how could this be?)
- a=median
- Quartiles=a +/- b
- Special case of Student's t with 1 d.f.
- Ratio of 2 independent unit normal variates is standard Cauchy variate
- Should not be thought of as "only a pathological case". (Casella & Berger) as we frequently (when?) calculate ratios of random variables.

Snedecor-Fisher's *F*-Distribution

Consider two independent random samples:

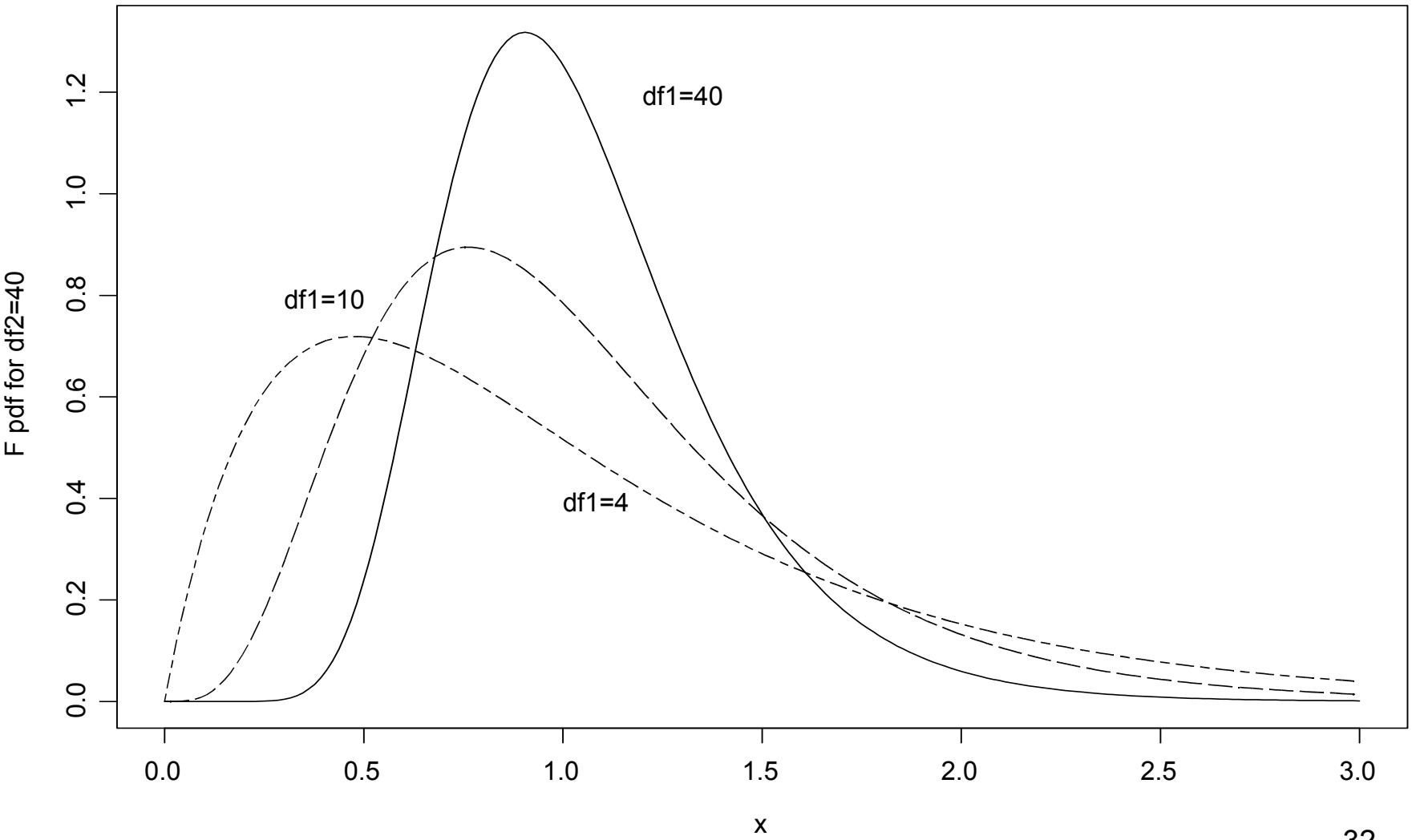
$X_1, X_2, ..., X_{n_1}$ from $N(\mu_1,\sigma_1^2)$, $Y_1, Y_2, ..., Y_{n_2}$ from $N(\mu_2,\sigma_2^2)$.

Then $$\frac{S_1^2/\sigma_1^2}{S_2^2/\sigma_2^2} = \frac{\left[\frac{(n_1-1)S_1^2}{\sigma_1^2}\right]/(n_1-1)}{\left[\frac{(n_2-1)S_2^2}{\sigma_2^2}\right]/(n_2-1)}$$

has an *F*-distribution with n_1-1 d.f. in the numerator and n_2-1 d.f. in the denominator.

- F is the ratio of two independent χ^2's divided by their respective d.f.'s
- Used to compare sample variances.
- See Table A.6, F-distribution, pp. 677-679

Snedecor's F Distribution



Snedecor's F Distribution Interesting Facts

- Parameters, v, w, referred to as degrees of freedom (df).
- Mean = w/(w-2), for w>2
- Variance = $2w^2(v+w-2)/(v(w-2)^2(w-4))$, for w>4
- As d.f., v and w increase, F variate tends to normal
- Related also to Chi-square, Student's t, Beta and Binomial
- Reference for distributions:

  Statistical Distributions 3rd ed. by Evans, Hastings and Peacock, Wiley, 2000

Sampling Distributions - Summary

- For random sample from **any distribution**, standardized sample mean converges to N(0,1) as n increases (CLT).
- In normal case, standardized sample mean with S instead of sigma in the denominator ~ Student's t(n-1).
- Sum of n squared unit normal variates ~ Chi-square (n)
- In the normal case, sample variance has scaled Chi-square distribution.
- In the normal case, ratio of sample variances from two different samples divided by their respective d.f. has F distribution.

Sir Ronald A. Fisher (1890-1962)

Wrote the first books on statistical methods (1926 & 1936):
"A student should not be made to read Fisher's books unless he has read them before."

George W. Snedecor (1882-1974)

Taught at Iowa State Univ. where wrote a college textbook (1937):
"Thank God for Snedecor; *now* we can understand Fisher."
(named the distribution for **F**isher)

Sampling Distributions for Order Statistics

Most sampling distribution results (except for CLT) apply to samples from normal populations.

If data does not come from a normal (or at least approximately normal), then statistical methods called “distribution-free” or “non-parametric” methods can be used (Chapter 14).

Non-parametric methods are often based on ordered data (called order statistics: $X_{(1)}, X_{(2)}, \ldots, X_{(n)}$) or just their ranks.

If $X_1..X_n$ are from a continuous population with cdf F(x) and pdf f(x) then the pdf of $X_{(j)}$ is:

$$f_{(j)}(x) = \frac{n!}{(j-1)!(n-j)!} f(x)[F(x)]^{j-1}[1-F(x)]^{n-j}$$

The confidence intervals for percentiles can be derived using the order statistics and the binomial distribution.