Chapter 1: A Review of Basic Concepts A

Dr. Daniel B. Rowe
Professor of Computational Statistics
Department of Mathematical and Statistical Sciences
Marquette University

A Review of Basic Concepts

Describing Quantitative Data Numerically

The mean of a sample of n measurements $y_1,\ldots,y_n$ is

$$\bar{y} = \frac{1}{n}\sum_{i=1}^{n} y_i$$

The mean of a population is $E(y)=\mu$.

The variance of a sample of n measurements $y_1,\ldots,y_n$ is

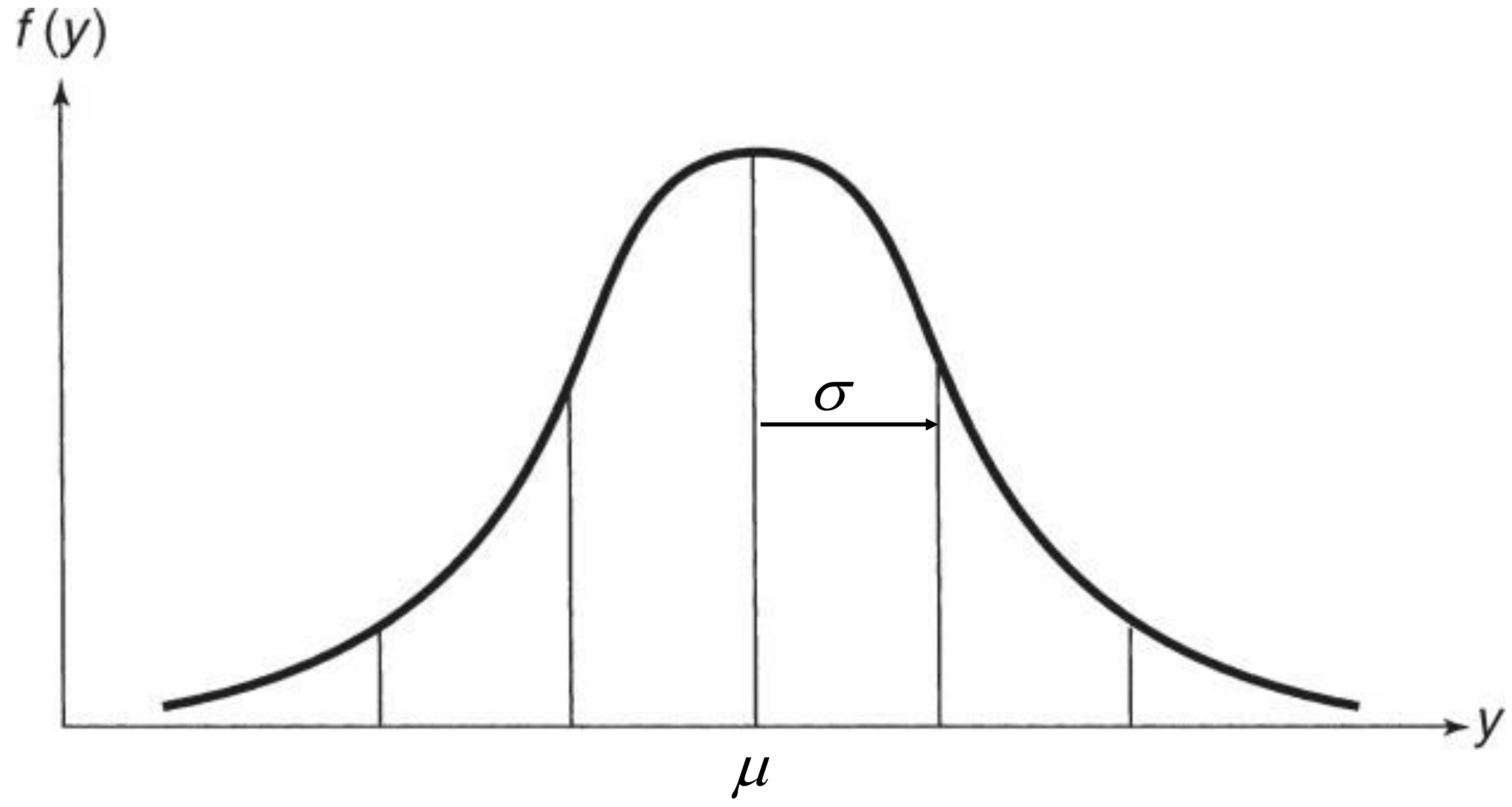
$$s^2 = \frac{1}{n-1}\sum_{i=1}^{n}(y_i-\bar{y})^2$$

The mean of a population is $E[(y-\mu)^2]=\sigma^2$.

The sample standard deviation is s and the population standard deviation is σ.

A Review of Basic Concepts

The Normal Probability Distribution



$$f(y\,|\,\mu,\sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

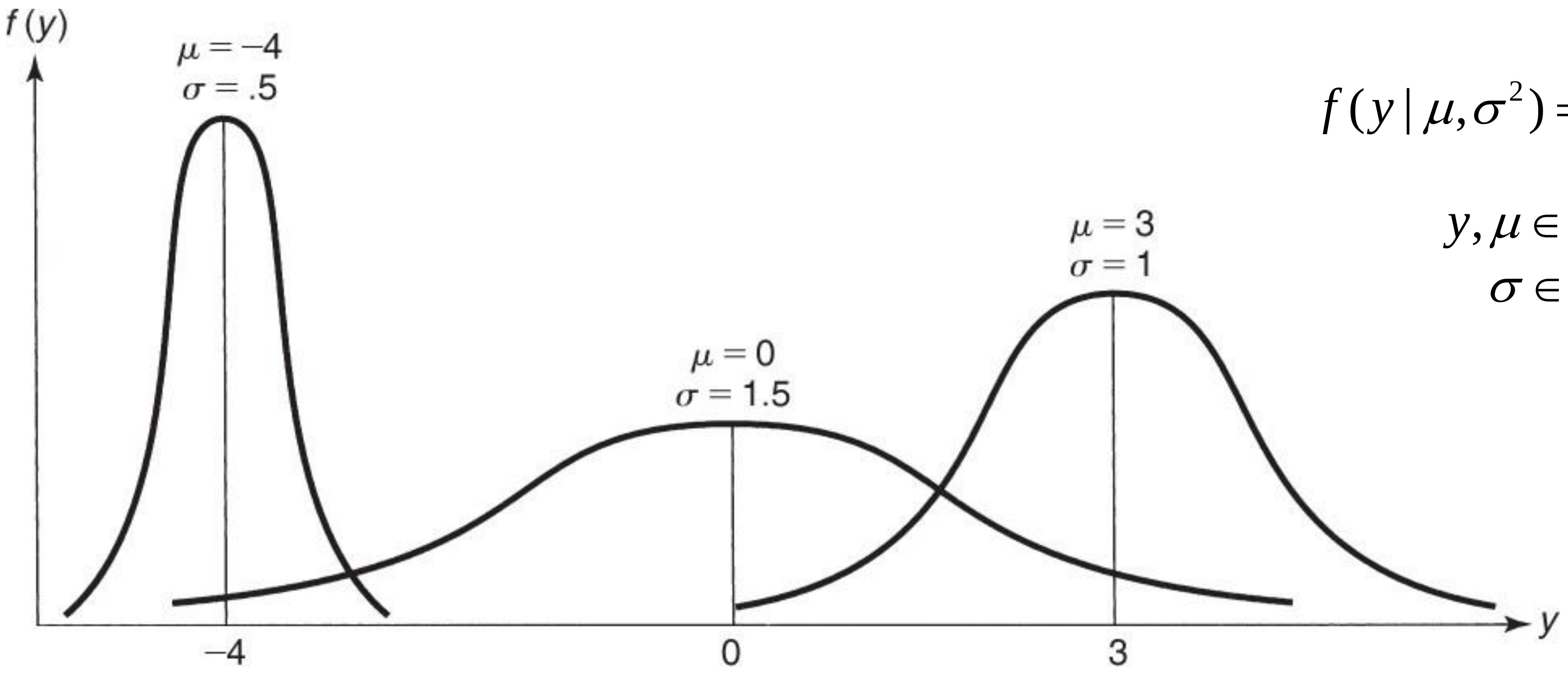
$$y, \mu \in \mathbb{R}$$

$$\sigma \in \mathbb{R}^+$$

e = 2.718281828459046…
π = 3.141592653589793…
μ = population mean
σ = population std. deviation

A Review of Basic Concepts

The Normal Probability Distribution



$$f(y\,|\,\mu,\sigma^2)=\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(y-\mu)^2}{2\sigma^2}}$$

$$y,\mu\in\mathbb{R}$$

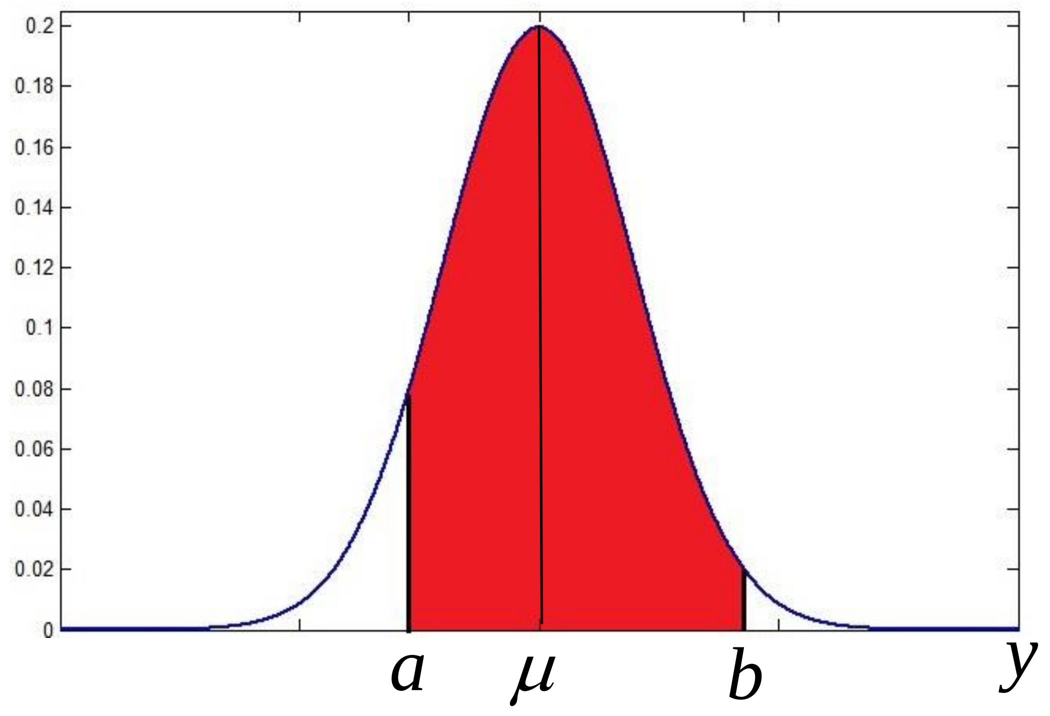
$$\sigma\in\mathbb{R}^+$$

A Review of Basic Concepts

The Normal Probability Distribution

Areas of continuous functions are found with Calculus.

$$A = \int_a^b \underbrace{\frac{1}{\sigma\sqrt{2\pi}} e^{-\frac{1}{2}\left(\frac{y-\mu}{\sigma}\right)^2}}_{f(y)} dy = P(a < y < b)$$



But we can’t integrate the normal distribution. So, we transform to standard normal.

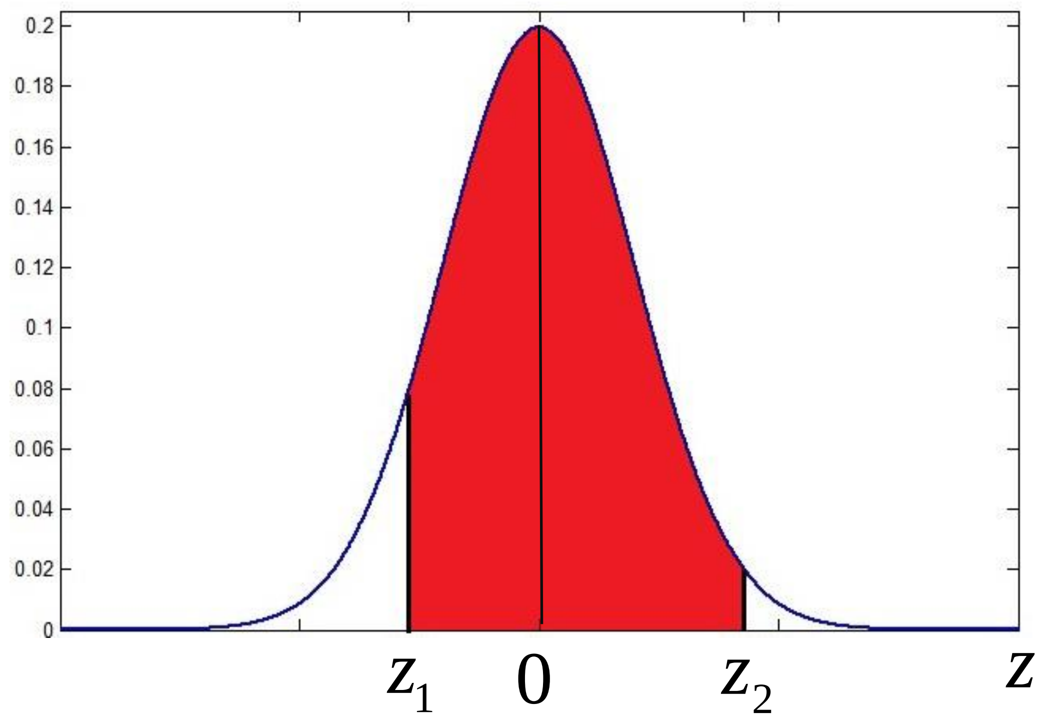
$$z = \frac{y-\mu}{\sigma}$$

And look up the areas in a table.

A Review of Basic Concepts

The Normal Probability Distribution

Convert a and b to z_1 and z_2.



$$z_1 = \frac{a-\mu}{\sigma}$$

$$z_2 = \frac{b-\mu}{\sigma}$$

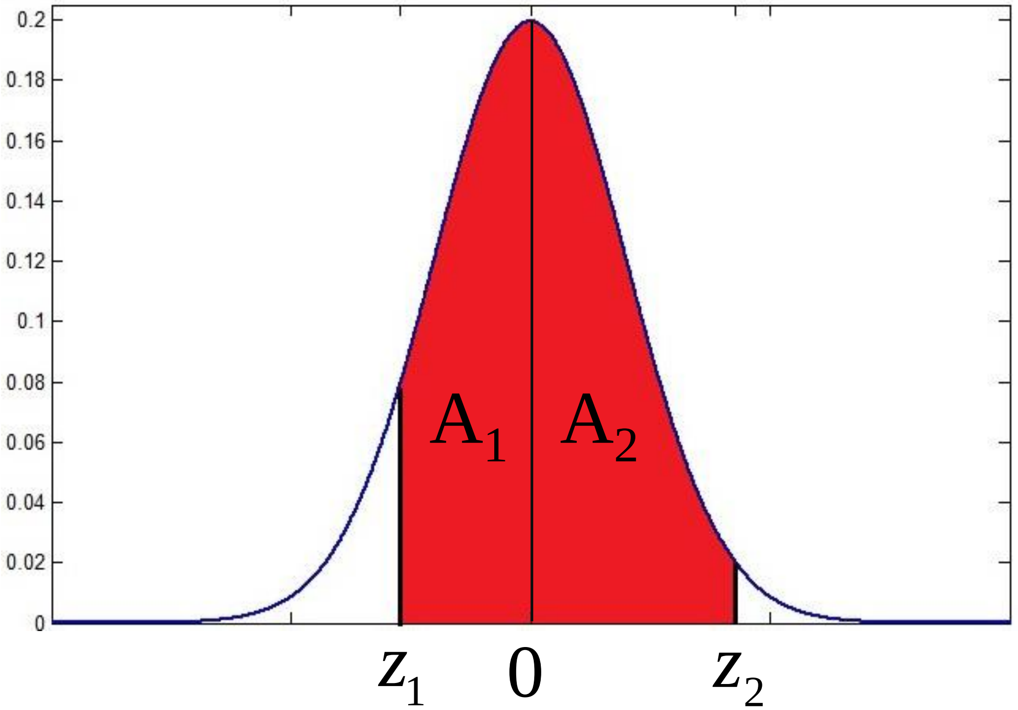
Look up area between z_1 and 0 as 0 to z_1.

Look up area between 0 and z_2.

Add together.

A Review of Basic Concepts

The Normal Probability Distribution



Look up area between z_1 and 0 as 0 to z_1.

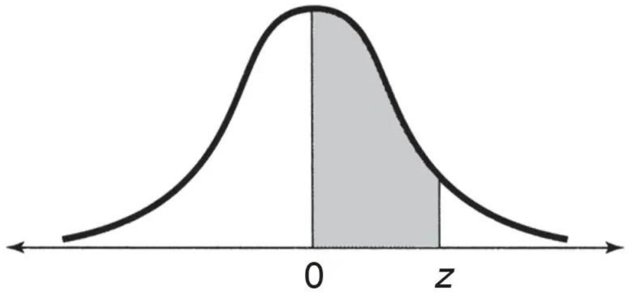
$A_1 = P(a<y<0)$

Look up area between 0 and z_2.

$A_2 = P(0<y<b)$

Add together.

$A_1 + A_2$



R Code

```
df <- 4
pval <- 0.975
qt(pval, df = df,
lower.tail = FALSE)
```

```
mean <- 0
sd <- 1
y <- 1.96
1-pnorm(y, mu, sd)
```

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0	.0000	.0040	.0080	.0120	.0160	.0199	.0239	.0279	.0319	.0359
.1	.0398	.0438	.0478	.0517	.0557	.0596	.0636	.0675	.0714	.0753
.2	.0793	.0832	.0871	.0910	.0948	.0987	.1026	.1064	.1103	.1141
.3	.1179	.1217	.1255	.1293	.1331	.1368	.1406	.1443	.1480	.1517
.4	.1554	.1591	.1628	.1664	.1700	.1736	.1772	.1808	.1844	.1879
.5	.1915	.1950	.1985	.2019	.2054	.2088	.2123	.2157	.2190	.2224
.6	.2257	.2291	.2324	.2357	.2389	.2422	.2454	.2486	.2517	.2549
.7	.2580	.2611	.2642	.2673	.2704	.2734	.2764	.2794	.2823	.2852
.8	.2881	.2910	.2939	.2967	.2995	.3023	.3051	.3078	.3106	.3133
.9	.3159	.3186	.3212	.3238	.3264	.3289	.3315	.3340	.3365	.3389
1.0	.3413	.3438	.3461	.3485	.3508	.3531	.3554	.3577	.3599	.3621
1.1	.3643	.3665	.3686	.3708	.3729	.3749	.3770	.3790	.3810	.3830
1.2	.3849	.3869	.3888	.3907	.3925	.3944	.3962	.3980	.3997	.4015
1.3	.4032	.4049	.4066	.4082	.4099	.4115	.4131	.4147	.4162	.4177
1.4	.4192	.4207	.4222	.4236	.4251	.4265	.4279	.4292	.4306	.4319
1.5	.4332	.4345	.4357	.4370	.4382	.4394	.4406	.4418	.4429	.4441

A Review of Basic Concepts

Sampling Distribution and the Central Limit Theorem

Theorem 1.1: Sampling distribution of the Sampling Mean

If $y_1,\ldots,y_n$ represent a random sample of n measurements ~~from a large (or infinite) population~~ with mean μ and standard deviation σ then, regardless of the form of the population relative distribution, the mean and standard error of estimate of the sampling distribution of $\bar{y}$ will be

Mean: $\mu_{\bar{y}} = E(y) = \mu$

Standard error of estimate: $\sigma_{\bar{y}} = \sqrt{E[(y-\mu)^2]} = \dfrac{\sigma}{\sqrt{n}}$

$$\mu = \int y f(y) dy < \infty$$

$$\sigma^2 = \int (y-\mu)^2 f(y) dy < \infty$$

A Review of Basic Concepts

Sampling Distribution and the Central Limit Theorem

Theorem 1.2: Central Limit Theorem

For large sample sizes, the mean $\bar{y}$ of a sample from a population with mean μ and standard deviation σ has a sampling distribution that is approximately normal, regardless of the probability distribution of the sampled population.
The larger the sample size, the better will be the normal approximation to the sampling distribution of $\bar{y}$.

A Review of Basic Concepts

Sampling Distribution and the CLT

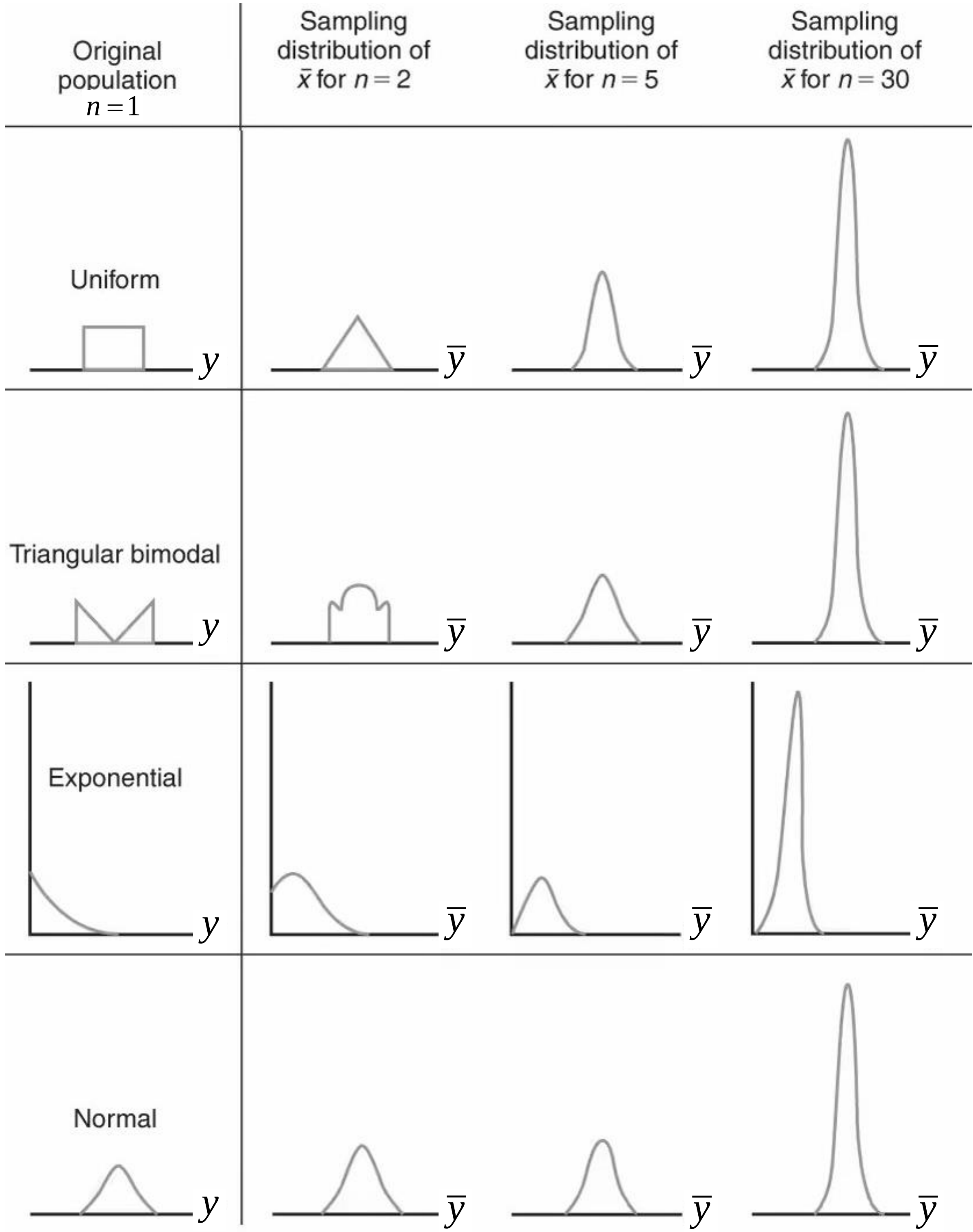
When n is large,

$$\bar{y} \sim N\left(\mu_{\bar{y}} = \mu, \sigma_{\bar{y}}^2 = \frac{\sigma^2}{n} \right) ,$$

no matter what distribution our original measurements come from, when n is large, $\bar{y}$ has a normal distribution

This means that we can use the z table to get areas!

$$z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$



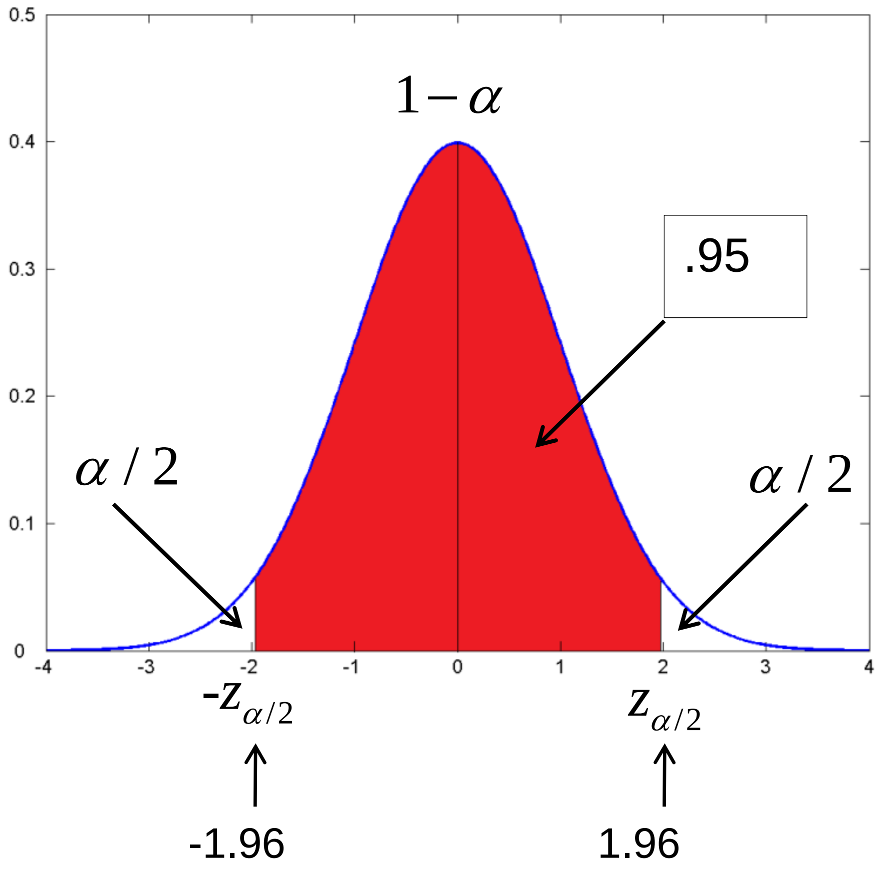
A Review of Basic Concepts

Estimating a Population Mean

When we estimate a parameter like μ with a single value like $\bar{y}$,
it is called a point estimator. We often are interested in a range of values within which we have a prespecified level of confidence that the interval contains μ.

We know that $P(-1.96<z<1.96)=0.95$,
or more generally, $P(-z_{\alpha/2}<z< z_{\alpha/2})=1-\alpha$.

Where $z_{\alpha/2}$ is called the confidence coefficient.
$z_{\alpha/2}$ is the value of z with an area $\alpha/2$ larger than it.



A Review of Basic Concepts

$$z = \frac{\bar{y} - \mu}{\sigma / \sqrt{n}}$$

Estimating a Population Mean

With some algebra on $P(-z_{\alpha/2} < z < z_{\alpha/2}) = 1 - \alpha$, we can see that …

$$z < z_{\alpha/2} \qquad\qquad -z_{\alpha/2} < z$$

$$\frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}} < z_{\alpha/2} \qquad\qquad -z_{\alpha/2} < \frac{\bar{y} - \mu_{\bar{y}}}{\sigma_{\bar{y}}}$$

$$\bar{y} - \mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}} \qquad\qquad -z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \bar{y} - \mu$$

$$-\mu < z_{\alpha/2} \frac{\sigma}{\sqrt{n}} - \bar{y} \qquad\qquad -z_{\alpha/2} \frac{\sigma}{\sqrt{n}} - \bar{y} < -\mu$$

$$\bar{y} - z_{\alpha/2} \frac{\sigma}{\sqrt{n}} < \mu \qquad\qquad \mu < \bar{y} + z_{\alpha/2} \frac{\sigma}{\sqrt{n}}$$

A Review of Basic Concepts

Estimating a Population Mean

Thus, a $(1\text{-}\alpha)\times 100\%$ confidence interval for μ is

$$\bar{y} - z_{\alpha/2}\frac{\sigma}{\sqrt{n}} \quad \text{to} \quad \bar{y} + z_{\alpha/2}\frac{\sigma}{\sqrt{n}}$$

which if α=0.05, a 95% confidence interval for μ is

$$\bar{y} - 1.96\frac{\sigma}{\sqrt{n}} \quad \text{to} \quad \bar{y} + 1.96\frac{\sigma}{\sqrt{n}} \ .$$

A Review of Basic Concepts

Estimating a Population Mean

However, we never know the true value of σ, so we replace it by s

$$\bar{y} - z_{\alpha/2}\frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{y} + z_{\alpha/2}\frac{s}{\sqrt{n}}$$

but then we also need to replace z by t, so our CI for μ is

$$\bar{y} - t_{\alpha/2,df}\frac{s}{\sqrt{n}} \quad \text{to} \quad \bar{y} + t_{\alpha/2,df}\frac{s}{\sqrt{n}} \;,$$

where df=n-1 is our degrees of freedom.

A Review of Basic Concepts

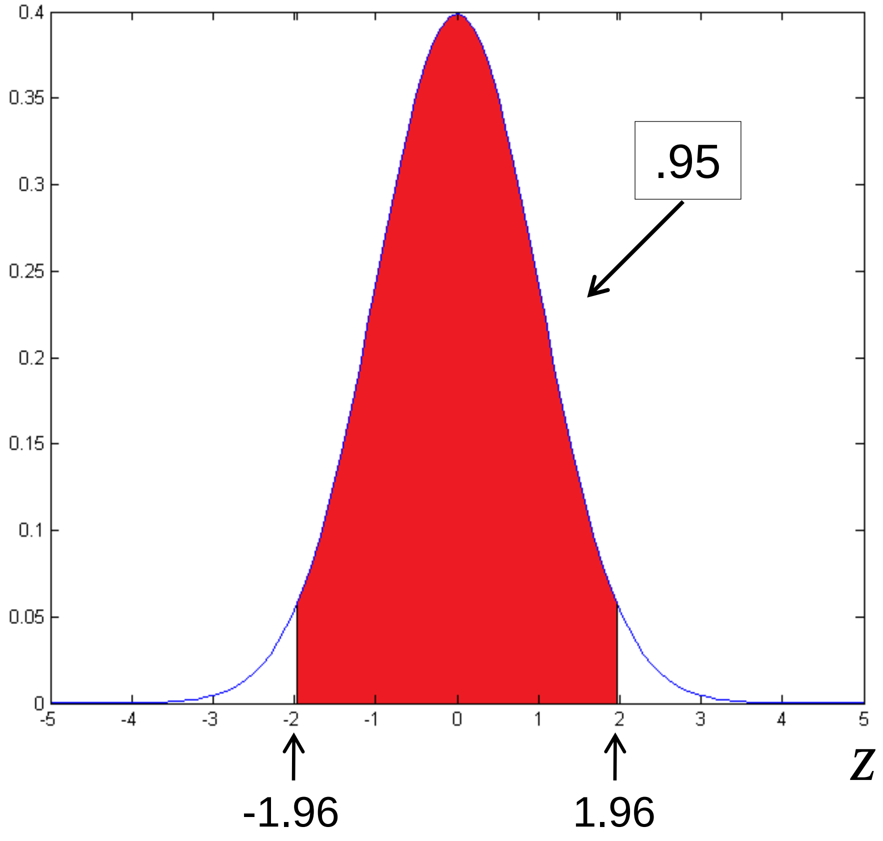
Estimating a Population Mean

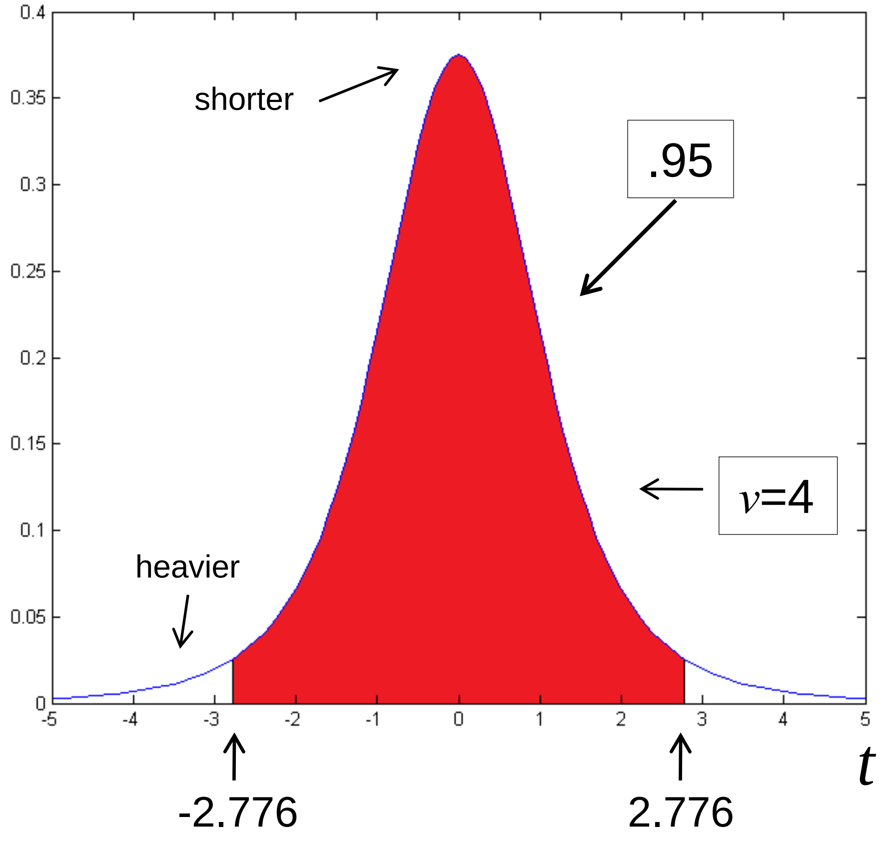
Since we estimated σ by s and changed z to t, the distribution and areas have changed.

$$f(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$

$$f(t|\nu) = \frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)} \frac{1}{\sqrt{\nu\pi}} \frac{1}{\left(1+\frac{1}{\nu}t^2\right)^{\frac{(\nu+1)}{2}}}$$

$$\nu = df = n - 1$$





A Review of Basic Concepts

Estimating a Population Mean

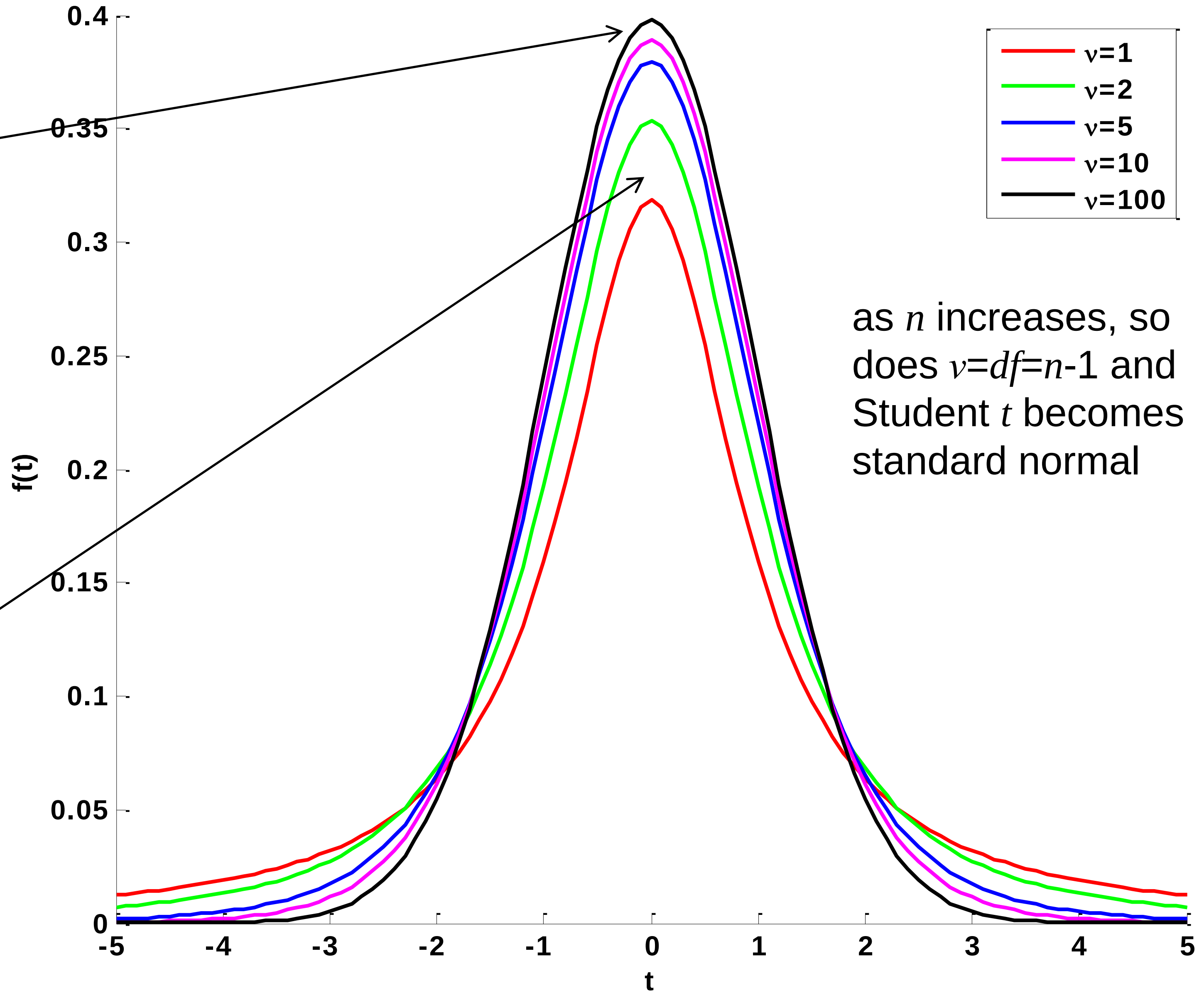
The standard normal dist. is:

$$f(z)=\frac{1}{\sqrt{2\pi}}e^{-\frac{1}{2}z^2}$$

The Student-t distribution is:

$$f(t|\nu)=\frac{\Gamma\left(\frac{\nu+1}{2}\right)}{\Gamma\left(\frac{\nu}{2}\right)}\frac{1}{\sqrt{\nu\pi}}\frac{1}{\left(1+\frac{1}{\nu}t^2\right)^{\frac{(\nu+1)}{2}}}$$

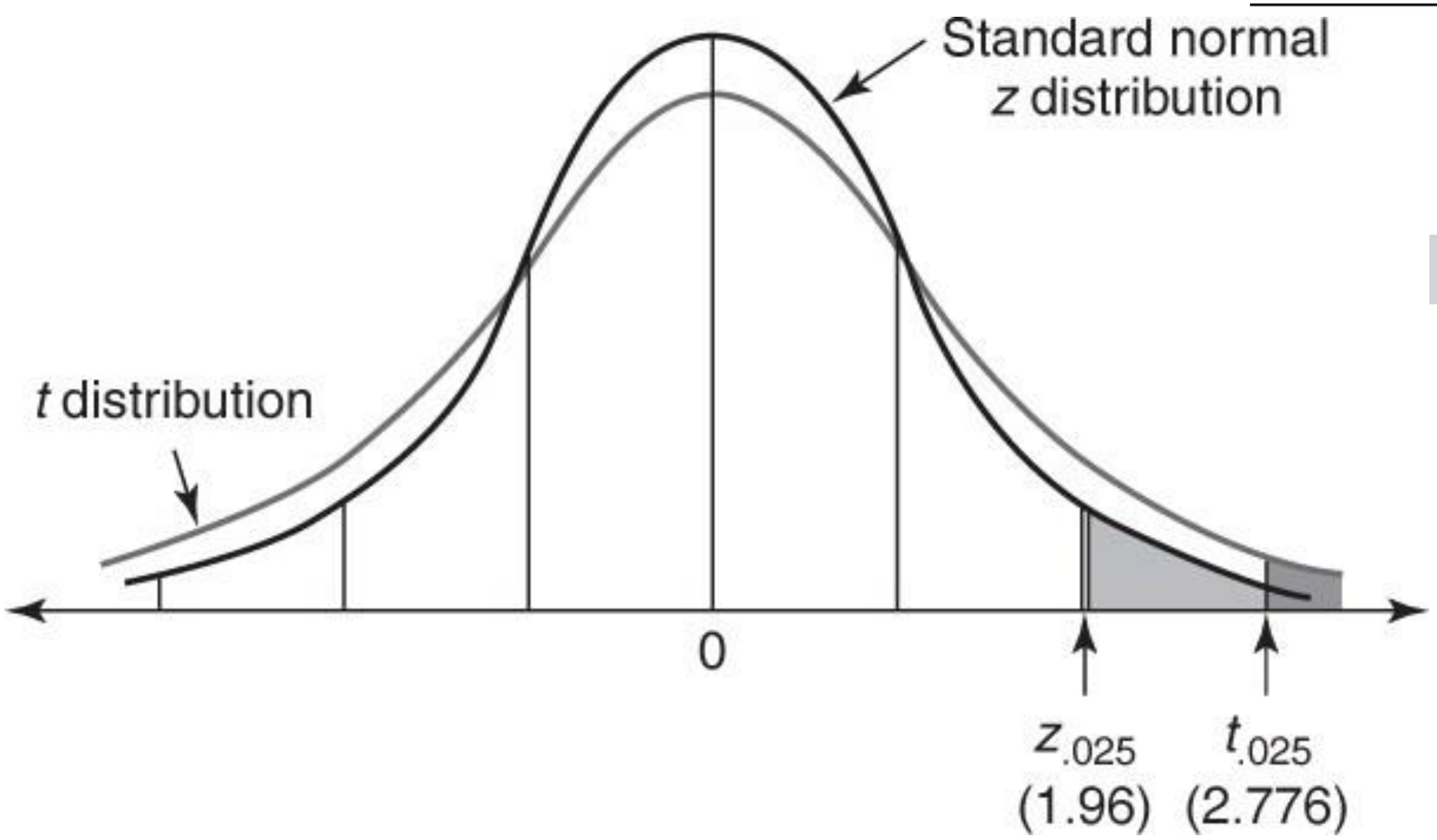
$$f(t|\nu)\rightarrow f(z)$$

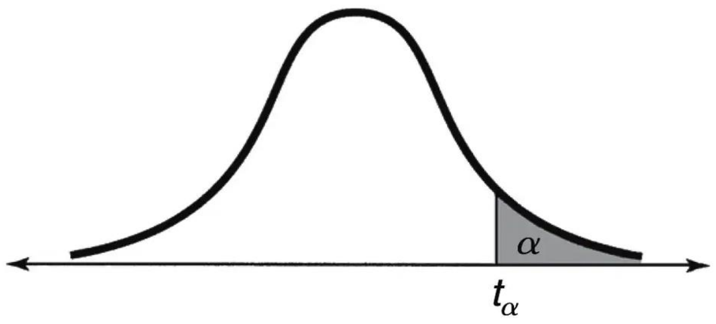


as n increases, so does $\nu=df=n$-1 and Student t becomes standard normal

A Review of Basic Concepts

Estimating a Population Mean





R Code

```
mean <- 0
sd <- 1
pval <- 0.975
qt(pval)
```

```
df <- 4
tval <- 2.776
pt(tval, df = df,
lower.tail = FALSE)
```

Degrees of Freedom	$t_{.100}$	$t_{.050}$	$t_{.025}$	$t_{.010}$	$t_{.005}$
1	3.078	6.314	12.706	31.821	63.657
2	1.886	2.920	4.303	6.965	9.925
3	1.638	2.353	3.182	4.541	5.841
4	1.533	2.132	2.776	3.747	4.604
5	1.476	2.015	2.571	3.365	4.032
6	1.440	1.943	2.447	3.143	3.707
7	1.415	1.895	2.365	2.998	3.499
8	1.397	1.860	2.306	2.896	3.355
9	1.383	1.833	2.262	2.821	3.250
10	1.372	1.812	2.228	2.764	3.169
11	1.363	1.796	2.201	2.718	3.106
12	1.356	1.782	2.179	2.681	3.055
13	1.350	1.771	2.160	2.650	3.012
14	1.345	1.761	2.145	2.624	2.977
15	1.341	1.753	2.131	2.602	2.947

A Review of Basic Concepts

Homework:

Read Chapter 1

A Review of Basic Concepts

Questions?