Distributions of Random Variables

In this Lecture we discuss the different types of random variables and illustrate the properties of typical probability distributions for these random variables.

What is a Random Variable?

A **variable** is any characteristic, observed or measured. A variable can be either **random** or **constant** in the population of interest.

Note this differs from common English usage where the word variable implies something that **varies** from individual to individual.

For a defined population, every **random variable** has an associated distribution that defines the **probability** of occurrence of each possible value of that variable (if there are a finitely countable number of unique values) or all possible sets of possible values (if the variable is defined on the real line).

Probability Distribution

A **probability distribution** (function) is a list of the probabilities of the values (simple outcomes) of a random variable.

Table: Number of heads in two tosses of a coin

y *outcome*	P(y) *probability*
0	1/4
1	2/4
2	1/4

For some experiments, the probability of a simple outcome can be easily calculated using a specific **probability function.** **I**f y is a simple outcome and p(y) is its probability.

$$0 \le p(y) \le 1$$

$$\sum_{\text{all y}} p(y) = 1$$

Discrete Distributions

Relative frequency distributions for "counting" experiments.

- Bernoulli Distribution
- Binomial Distribution
- Negative Binomial
- Poisson Distribution
- Geometric Distribution
- Multinomial Distribution

Yes-No responses.

Sums of Bernoulli responses

Number of trials to k^{th} event

Points in given space

Number of trials until first *success*

Multiple possible outcomes for each trial

Binomial Distribution

- The experiment consists of **n identical trials** (simple experiments).
- Each trial results in one of **two outcomes** (success or failure)
- The probability of success on a single trial is equal to π and π remains the same from trial to trial.
- The trials are independent, that is, the outcome of one trial does not influence the outcome of any other trial.
- The random variable y is the number of successes observed during n trials.

$$P(y) = \frac{n!}{y!(n-y)!} \pi^{y} (1-\pi)^{n-y}$$

n!=1x2x3x…x n

$$\boxed{\begin{array}{c} \mu = n\pi \\ \sigma = \sqrt{n\pi(1-\pi)} \end{array}}$$

Mean

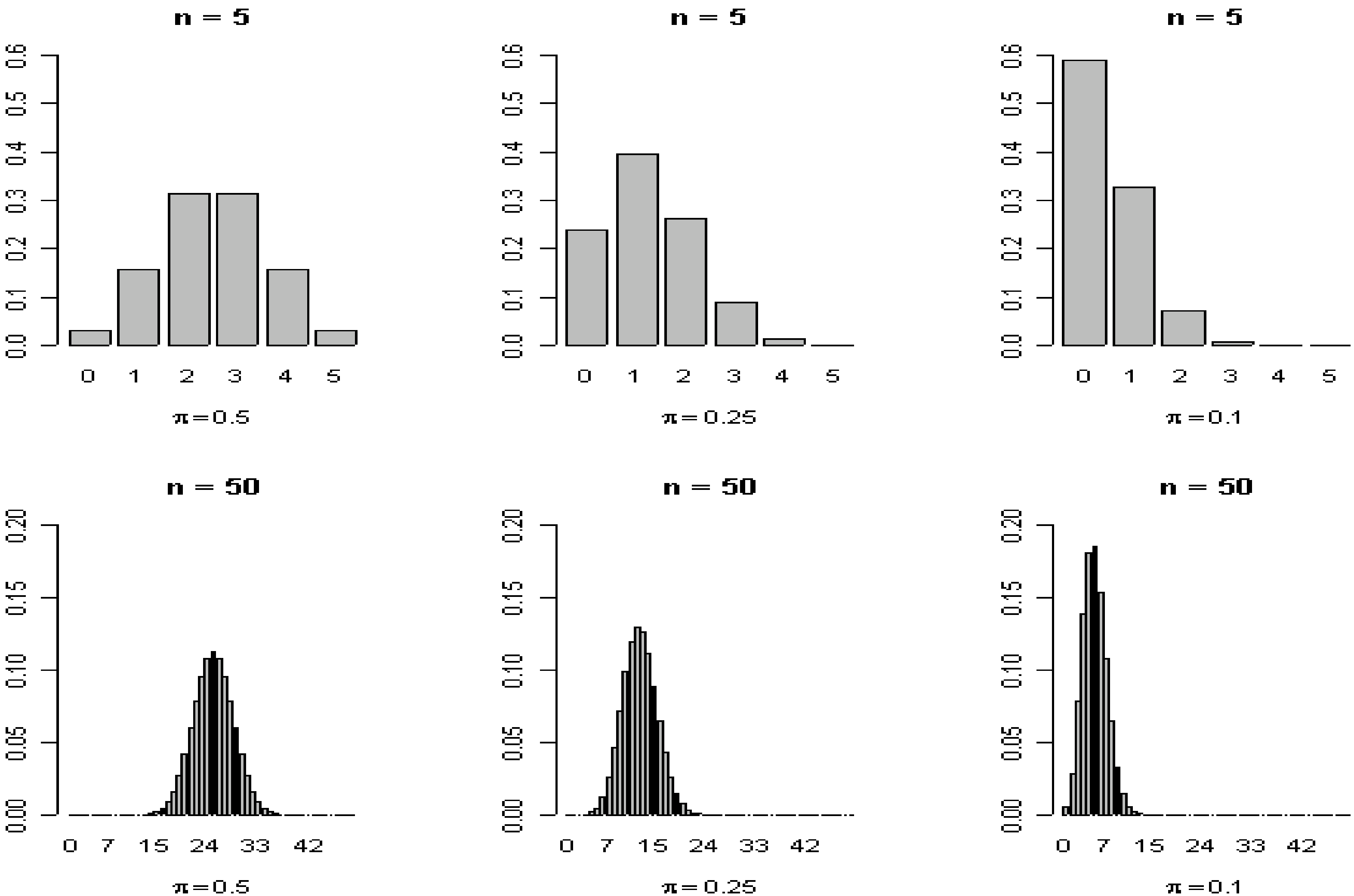
Standard deviation

Example n=5

$$P(y)=\frac{n!}{y!(n-y)!}\pi^{y}(1-\pi)^{n-y}$$

		n	π	π	π	π
		5	0.5	0.25	0.1	0.05
y	y!	n!/(y!)(n-y)!	P(y)	P(y)	P(y)	P(y)
0	1	1	0.03125	0.2373	0.59049	0.7737809
1	1	5	0.15625	0.3955	0.32805	0.2036266
2	2	10	0.31250	0.2637	0.07290	0.0214344
3	6	10	0.31250	0.0879	0.00810	0.0011281
4	24	5	0.15625	0.0146	0.00045	0.0000297
5	120	1	0.03125	0.0010	0.00001	0.0000003
		sum =	1	1	1	1

Binomial probability density function forms



As the n goes up, the distribution looks more symmetric and bell shaped.

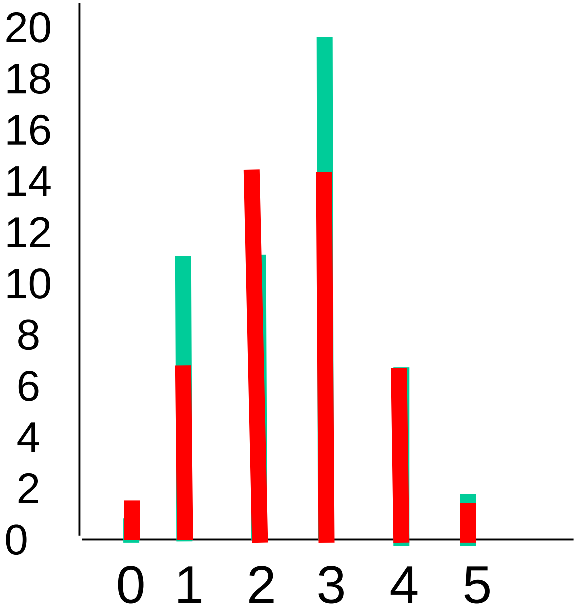
Binomial Distribution Example

Basic Experiment: 5 fair coins are tossed.
Event of interest: total number of heads.

Each coin is a trial with probability of a head coming up (a success) equal to 0.5. So the number of heads in the five coins is a binomial random variable with n=5 and π=.5.

The Experiment is repeated 50 times.

# of heads	*Observed*	*Theoretical*
0	1	1.56
1	11	7.81
2	11	15.63
3	19	15.63
4	6	7.81
5	2	1.56



Poisson Distribution

A random variable is said to have a *Poisson Distribution* with rate parameter λ, if its probability function is given by:

e=2.718…

$$P(y) = \frac{\lambda^y}{y!} e^{-\lambda}, \qquad \text{for } y = 0,1,2,\ldots$$

$$\boxed{\mu = \lambda, \quad \sigma^2 = \lambda}$$ ← Mean and variance for a Poisson

Ex: A certain type of tree has seedlings randomly dispersed in a large area, with a mean density of approx 5 per sq meter. If a 10 sq meter area is randomly sampled, what is the probability that no such seedlings are found?

$P(0) = 50^0(e^{-50})/0! =$ approx 10^{-22}

(Since this probability is so small, if no seedlings were actually found, we might question the validity of our model…)

Environmental Example

Van Beneden (1994,Env. Health. Persp., 102, Suppl. 12, p.81-83) describes an experiment where DNA is taken from softshell and hardshell clams and transfected into murine cells in culture in order to study the ability of the murine host cells to indicate selected damage to the clam DNA. (Mouse cells are much easier to culture than clam cells. This process could facilitate laboratory study of *in vivo* aquatic toxicity in clams).

The response is the **number** of focal lesions seen on the plates after a fixed period of incubation. The goal is to assess whether there are differences in response between DNA transfected from two clam species.

The response could be modeled as if it followed a **Poisson Distribution**.

Ref: Piegorsch and Bailer, Stat for Environmental Biol and Tox, p400

Discrete Distributions
Take Home Messages

- Primarily related to “counting” experiments.
- Probability only defined for “integer” values.
- Symmetric and non-symmetric distribution shapes.
- Best description is a frequency table.

Examples where discrete distributions are seen.

Wildlife - animal sampling, birds in a 2 km x 2 km area.
Botany - vegetation sampling, quadrats, flowers on stem.
Entomology - bugs on a leaf
Medicine - disease incidence, clinical trials
Engineering - quality control, number of failures in fixed time

Continuous Distributions

- **Normal Distribution**
- Log Normal Distribution
- Gamma Distribution
- **Chi Square Distribution**
- **F Distribution**
- **t Distribution**
- Weibull Distribution
- Extreme Value Distribution (Type I and II)

Foundations for much of statistical inference

Environmental variables

Time to failure, radioactivity

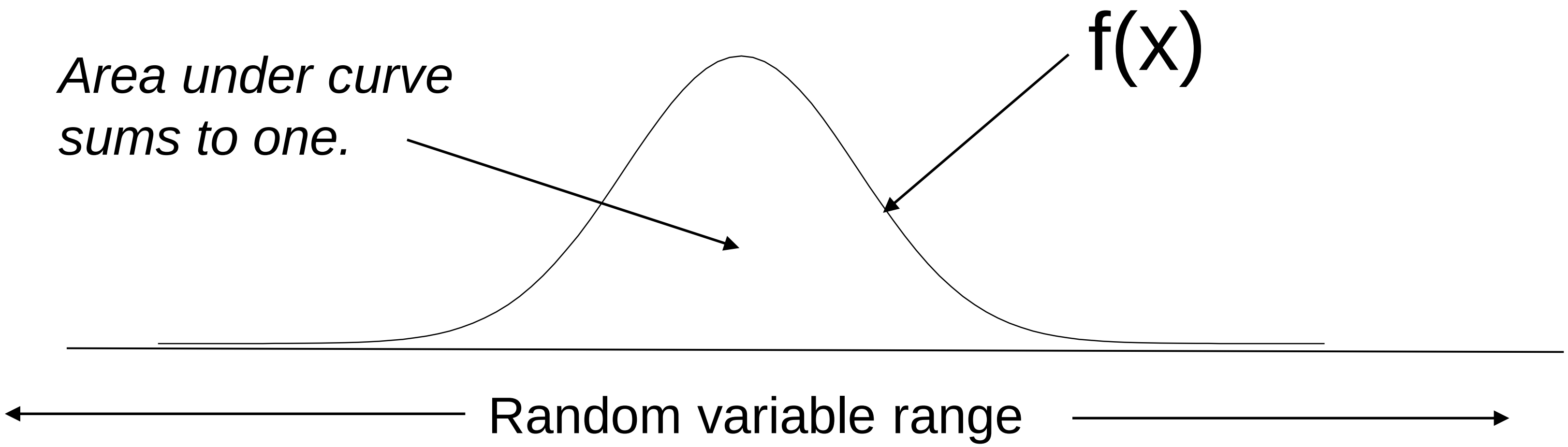
Basis for statistical tests.

Lifetime distributions

Continuous random variables are defined for continuous numbers on the real line. Probabilities have to be computed for all possible sets of numbers.

Probability Density Function

A function which integrates to 1 over its range and from which event probabilities can be determined.



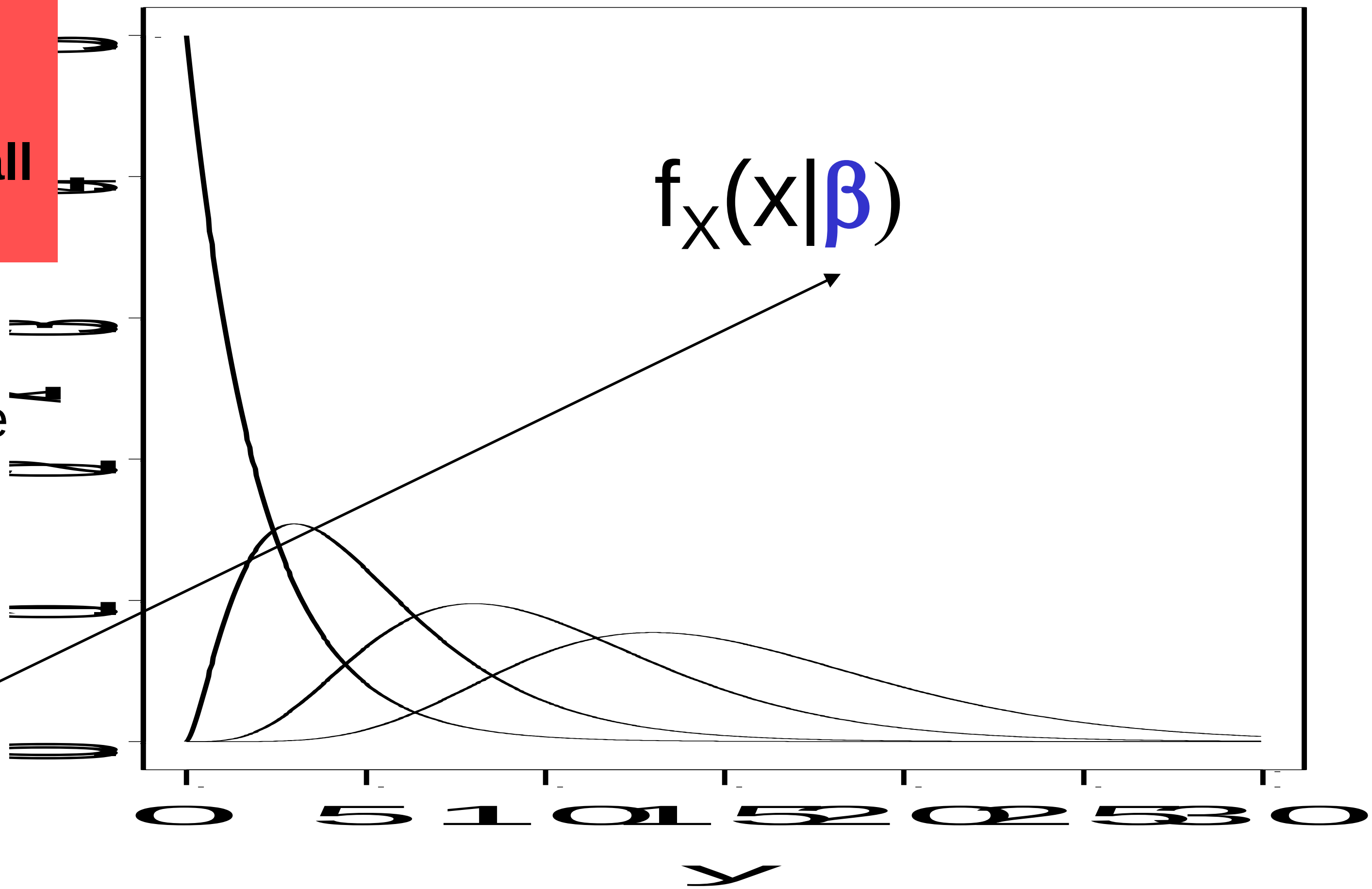
A theoretical shape - if we were able to sample the whole (infinite) population of possible values, this is what the associated histogram would look like.

A mathematical abstraction

Probability Density Function

The pdf does not have to be symmetric, nor be defined for all real numbers.

Chi Square density functions



The shape of the curve is determined by one or more **distribution parameters.**

Continuous Distribution Properties

Probability can be computed by integrating the density function.

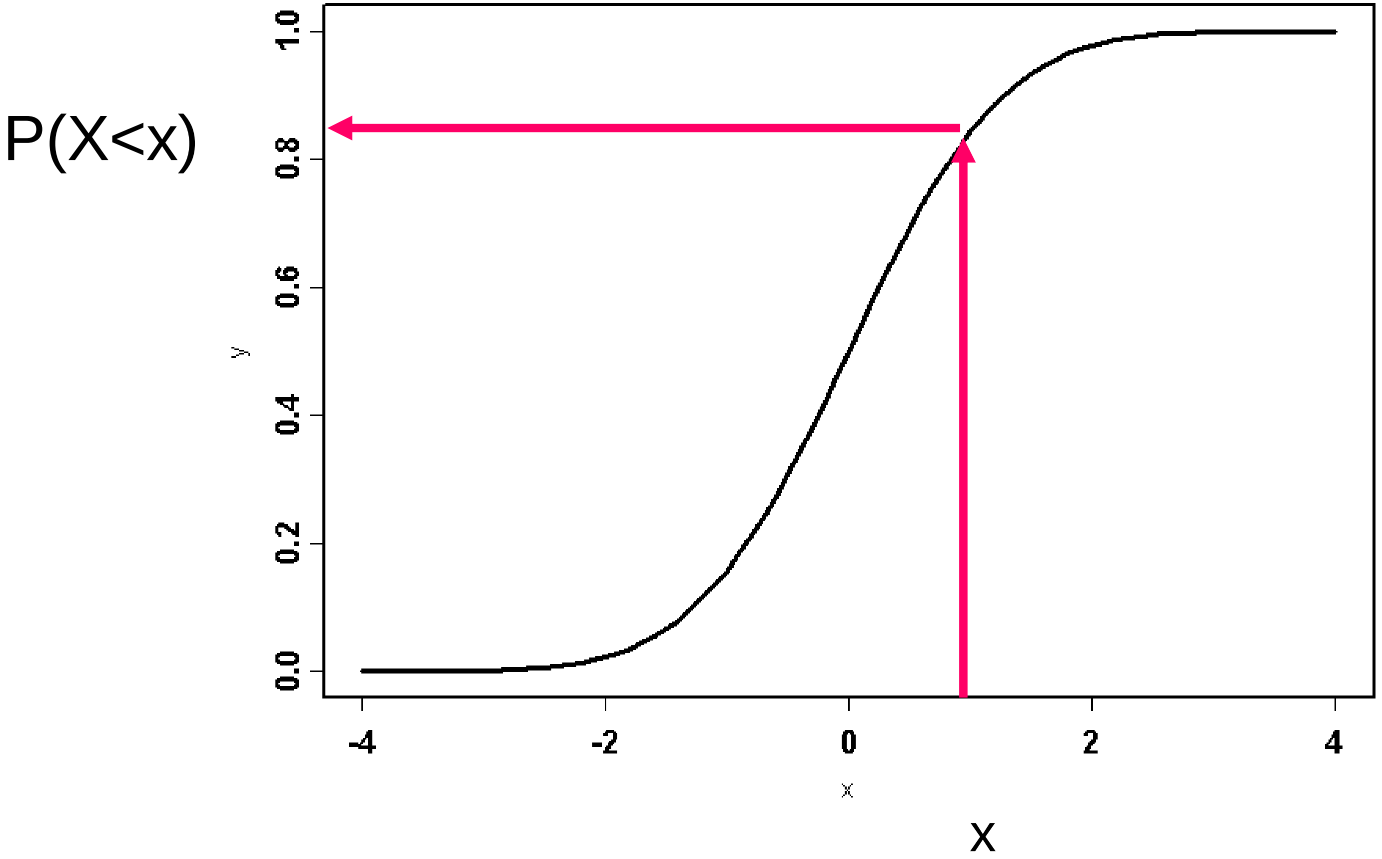
$$F(x_0) = P(X < x_0) = \int_{-\infty}^{x_o} f_X(x)dx$$

Continuous random variables only have positive probability for events which define intervals on the real line.

Any one point has **zero probability of occurrence**.

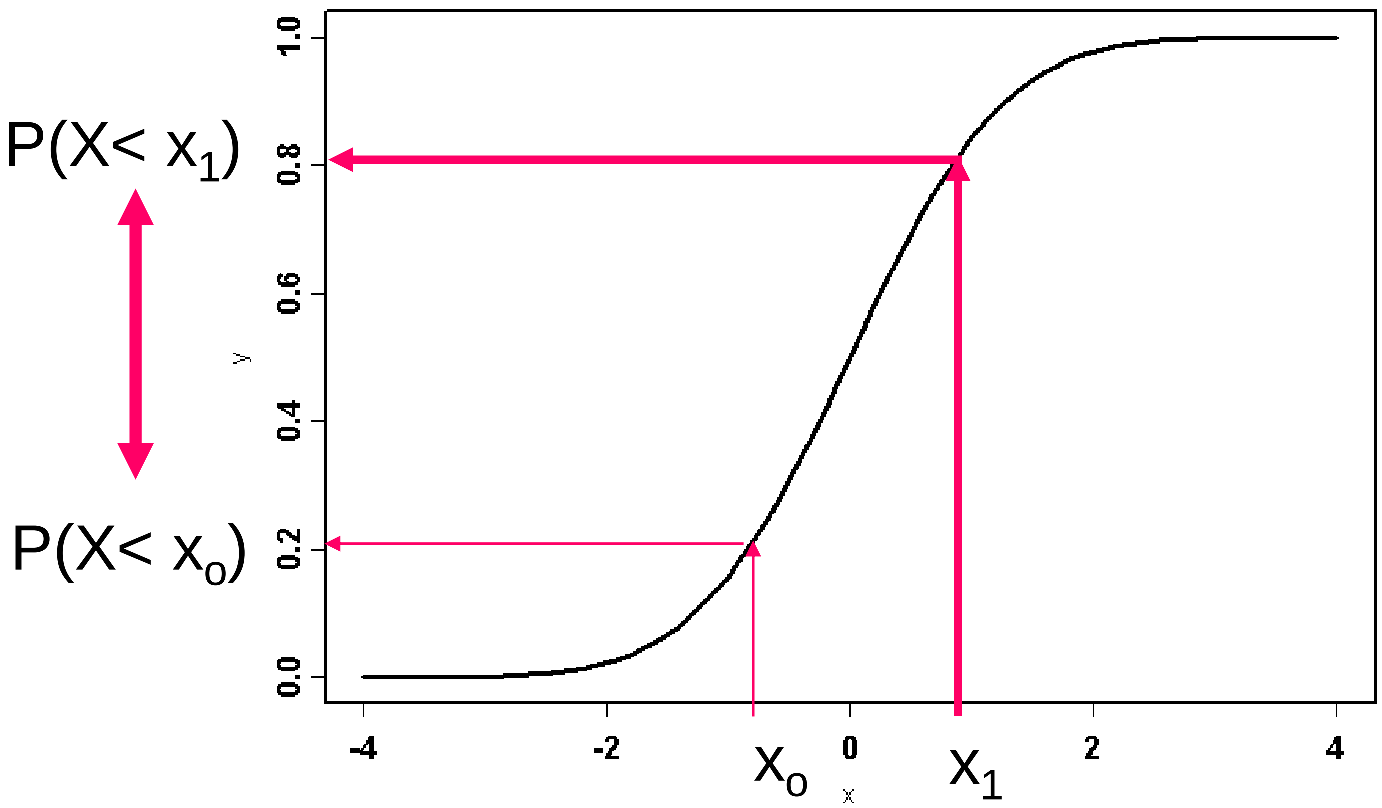
$$P(X = x_0) = \int_{x_0}^{x_0} f_X(x)dx = 0$$

Cumulative Distribution Function



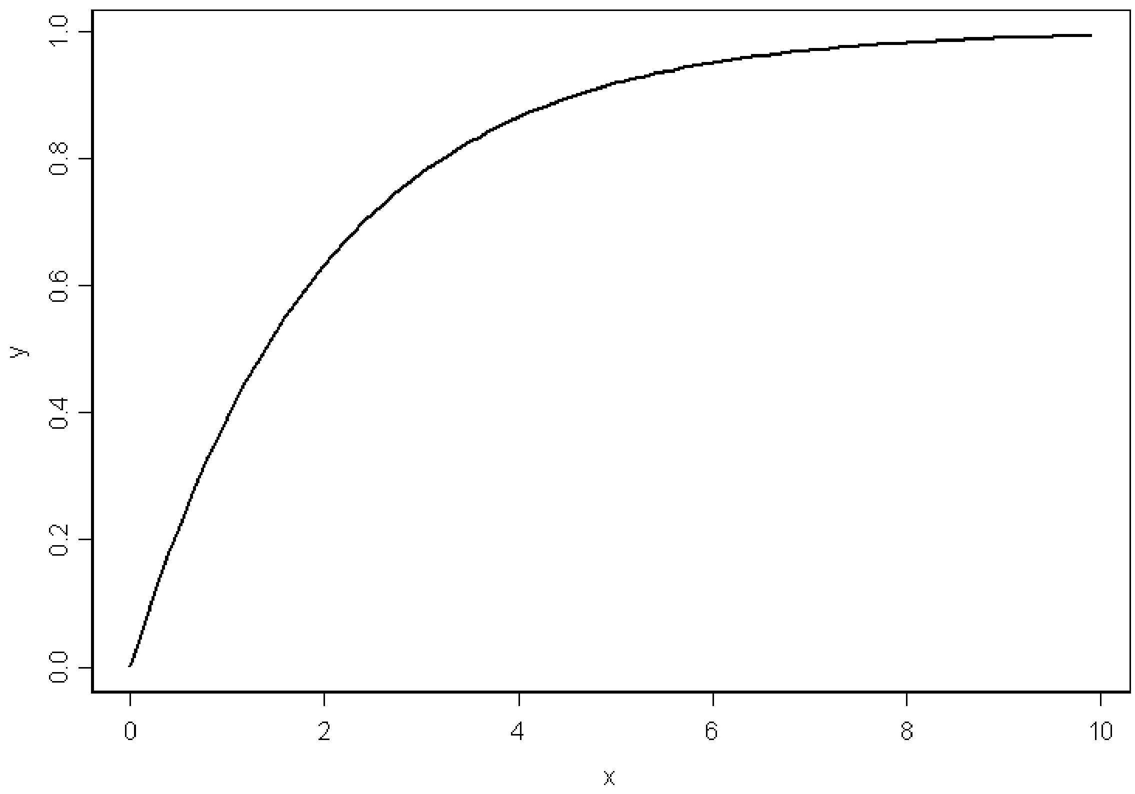
Using the Cumulative Distribution

$P(x_o < X < x_1) = P(X < x_1) - P(X < x_o) = .8 - .2 = .6$



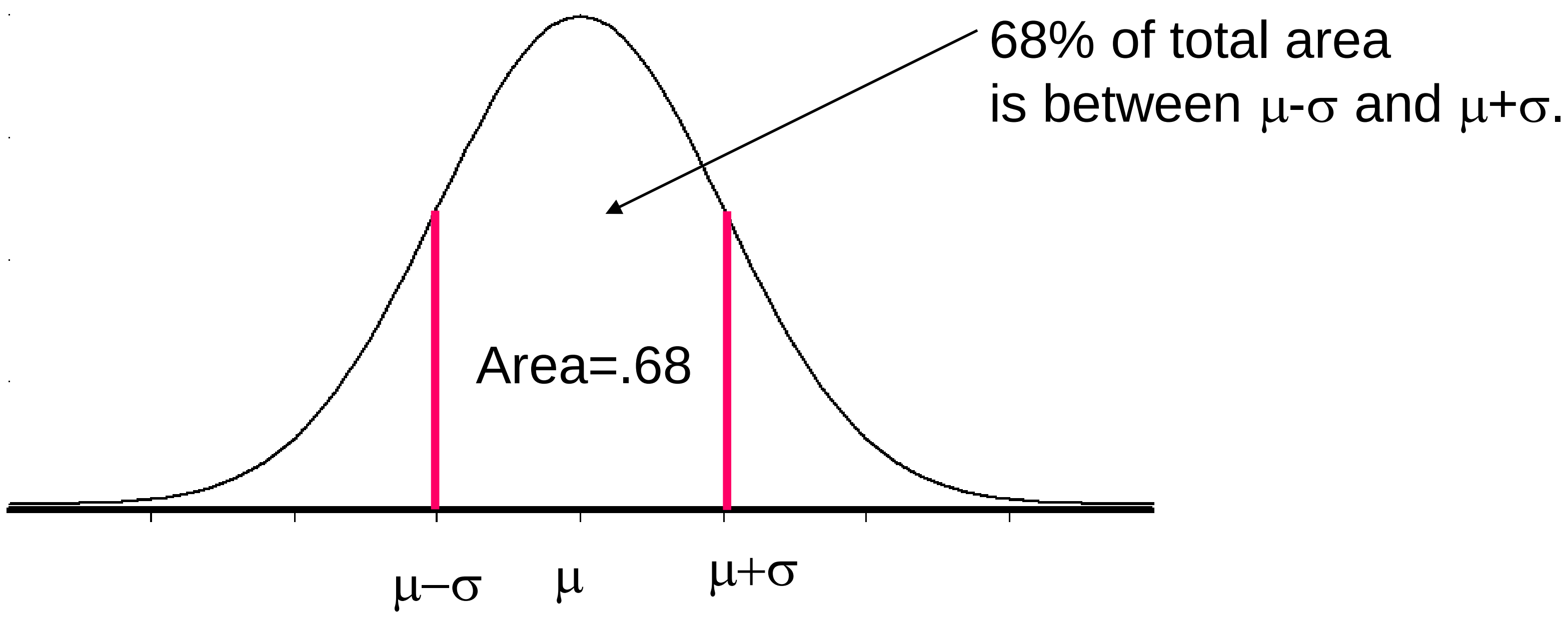
Chi Square Cumulative Distribution

Cumulative distribution does not have to be S shaped. In fact, only the normal and t-distributions have S shaped distributions.



Normal Distribution

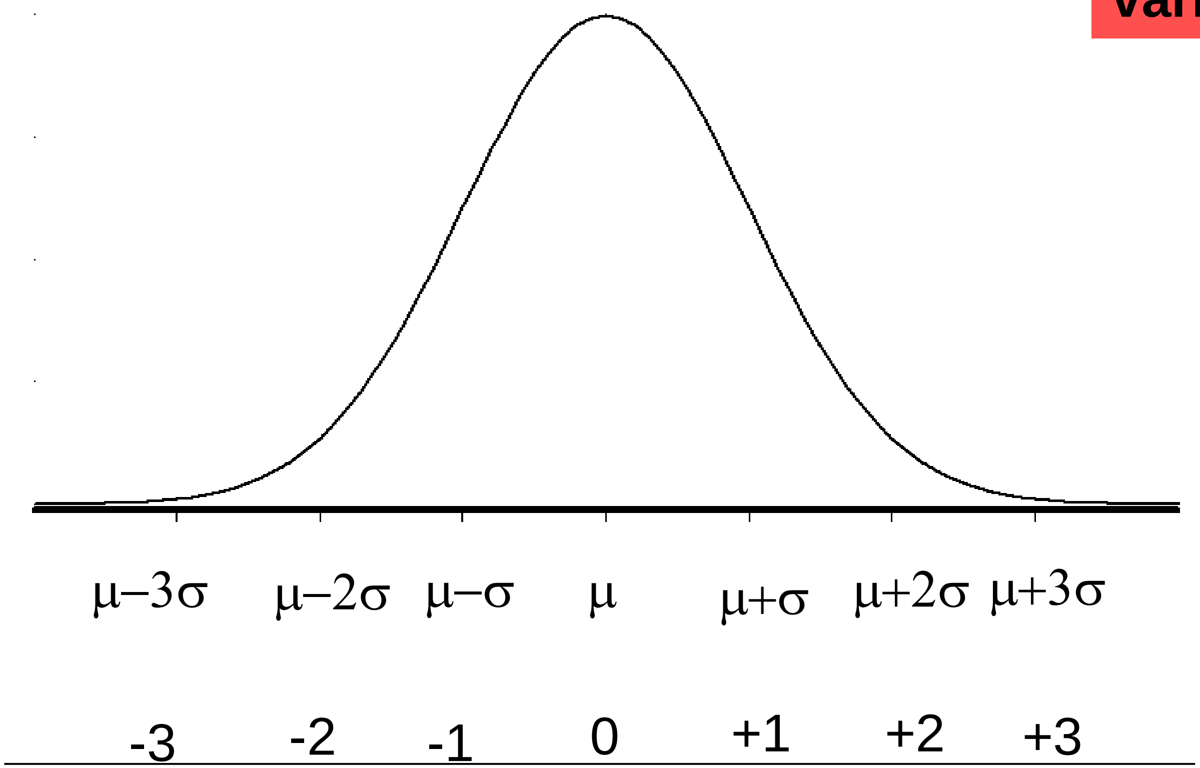
A symmetric distribution defined on the range $-\infty$ to $+\infty$ whose shape is defined by two parameters, the **mean**, denoted μ, that centers the distribution, and the **standard deviation**, σ, that determines the spread of the distribution.



$$P(\mu - \sigma < X < \mu + \sigma) = .68$$

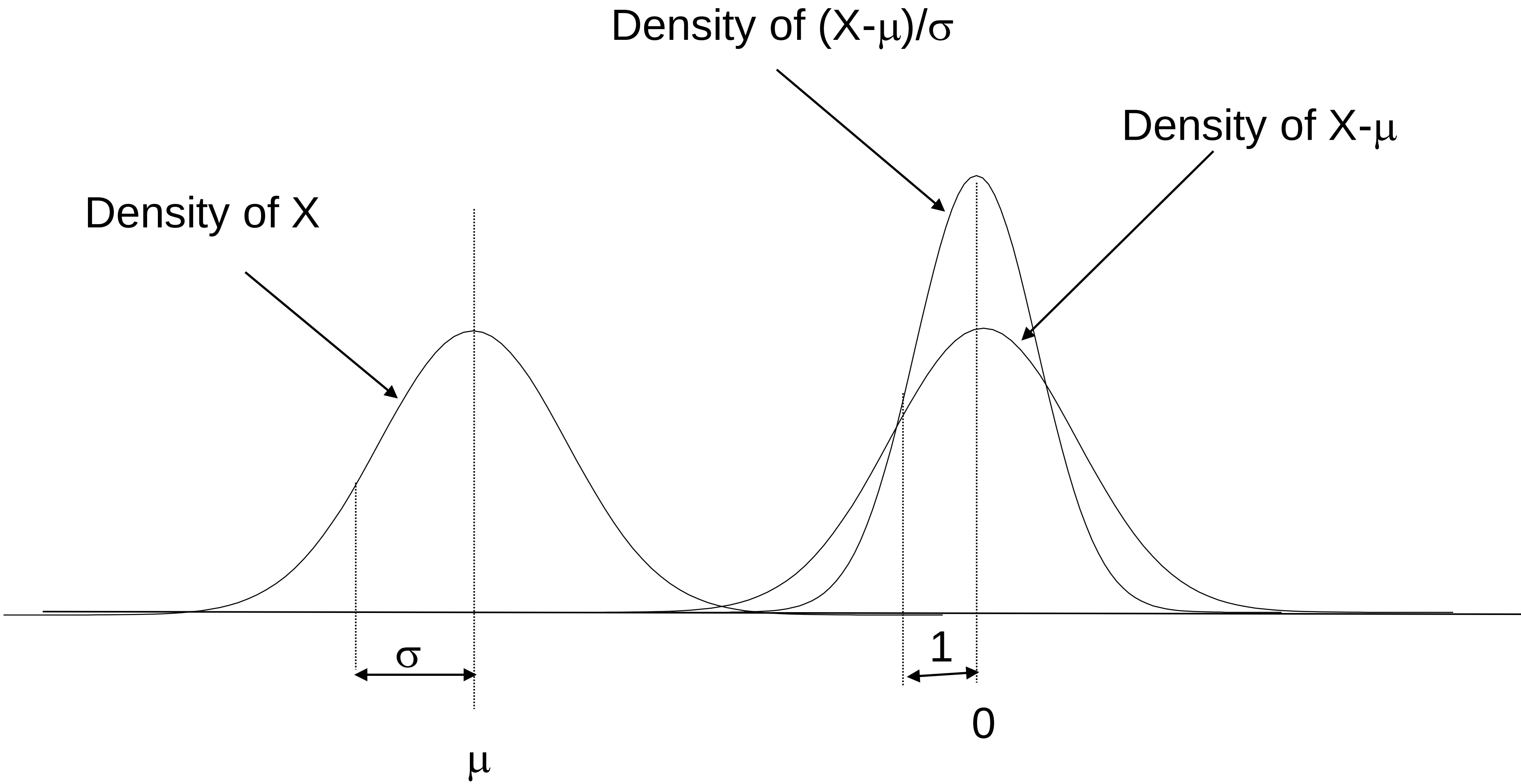
Standard Normal Distribution

All normal random variables can be related back to the **standard normal random variable**.



A Standard Normal random variable has mean 0 and standard deviation 1.

Illustration



Notation

Suppose **X** has a normal distribution with mean μ and standard deviation σ, denoted **X ~ N(μ, σ).**

Then a new random variable defined as **Z=(X- μ)/ σ**, has the standard normal distribution, denoted **Z ~ N(0,1)**.

Z=(X- μ)/ σ

To **standardize** we subtract the mean and divide by the standard deviation.

σ Z+ μ = X

To create a random variable with specific mean and standard deviation, we start with a standard normal deviate, multiply it by the target standard deviation, and then add the target mean.

Why is this important? Because in this way, the probability of any event on a normal random variable with any given mean and standard deviation can be computed from tables of the standard normal distribution.

Relating Any Normal RV to a Standard Normal RV

$$X \sim N(\mu, \sigma) \qquad Z \sim N(0,1)$$

$$X = \sigma Z + \mu$$

$$\begin{aligned} P(X < x_0) &= P(\sigma Z + \mu < x_0) \\ & P(\sigma Z < x_0 - \mu) \\ & P\left(Z < \boxed{\frac{x_0 - \mu}{\sigma}}\right) \end{aligned}$$

This probability can be found in a table of standard normal probabilities (Table 1 in Ott and Longnecker)

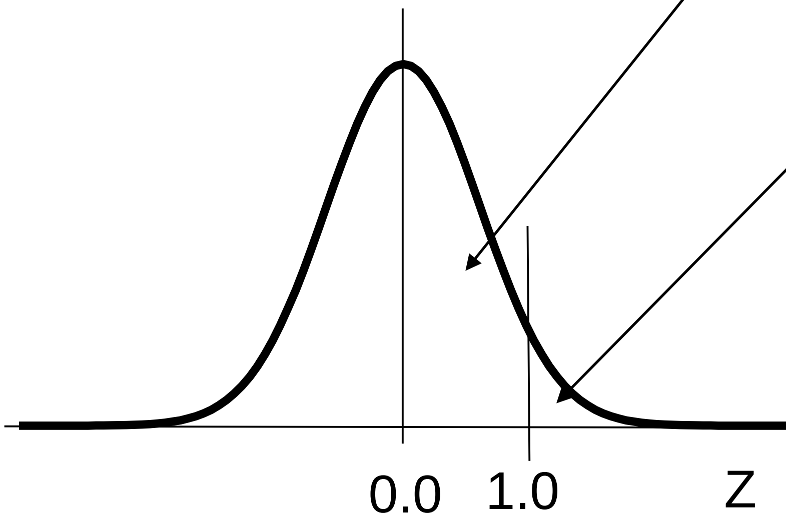
This value is just a number, usually between ± 4

Other Useful Relationships

$P(Z < z_0) = 1 - P(Z > z_0)$ ← Probability of complementary events.

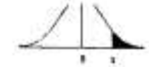
$P(Z < -z_0) = P(Z > +z_0)$ ← Symmetry of the normal distribution.

Normal Table



Ott & Longnecker, Table 1 page 676, gives areas left of z. This table from a previous edition gives areas right of z.

TABLE 2 Upper-tail Areas for the Normal Curve



z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	$\boxed{.5000}$	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.50	.3085	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.00	$\boxed{.1587}$	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.50	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.80	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.30	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.40	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.50	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.60	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.70	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.80	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.90	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.00	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

z	Area
3.500	.00023263
4.000	.00003167
4.500	.00000340
5.000	.00000029

Source: Computed by J. W. Stegeman using SAS.

Page A-3 OTT 1998
An Introduction To
STATISTICAL methods
and Data Analysis

Find P(2 < X < 4) when X ~ N(5,2).
The standardization equation for X is:

$$Z = (X-\mu)/\sigma = (X-5)/2$$

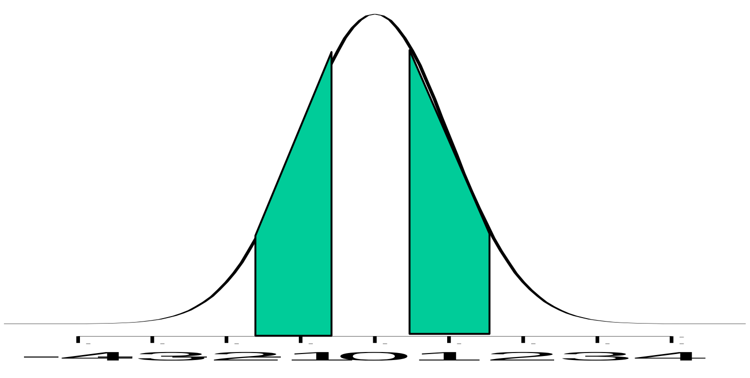
when X=2, Z= -3/2 = -1.5
when X=4, Z= -1/2 = -0.5

P(2<X<4) = P(X<4) - P(X<2)

P(X<2) = P(Z< -1.5)
= P(Z > 1.5) (by symmetry)

P(X<4) = P(Z < -0.5)
= P(Z > 0.5) (by symmetry)

P(2 < x < 4) = P(X<4)-P(X<2)
= P(Z>0.5) - P(Z > 1.5)
= 0.3085 - 0.0668 = 0.2417



Using a Normal Table

TABLE 2 Upper-tail Areas for the Normal Curve

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.00	.5000	.4960	.4920	.4880	.4840	.4801	.4761	.4721	.4681	.4641
0.10	.4602	.4562	.4522	.4483	.4443	.4404	.4364	.4325	.4286	.4247
0.20	.4207	.4168	.4129	.4090	.4052	.4013	.3974	.3936	.3897	.3859
0.30	.3821	.3783	.3745	.3707	.3669	.3632	.3594	.3557	.3520	.3483
0.40	.3446	.3409	.3372	.3336	.3300	.3264	.3228	.3192	.3156	.3121
0.50	$\boxed{.3085}$	.3050	.3015	.2981	.2946	.2912	.2877	.2843	.2810	.2776
0.60	.2743	.2709	.2676	.2643	.2611	.2578	.2546	.2514	.2483	.2451
0.70	.2420	.2389	.2358	.2327	.2296	.2266	.2236	.2206	.2177	.2148
0.80	.2119	.2090	.2061	.2033	.2005	.1977	.1949	.1922	.1894	.1867
0.90	.1841	.1814	.1788	.1762	.1736	.1711	.1685	.1660	.1635	.1611
1.00	.1587	.1562	.1539	.1515	.1492	.1469	.1446	.1423	.1401	.1379
1.10	.1357	.1335	.1314	.1292	.1271	.1251	.1230	.1210	.1190	.1170
1.20	.1151	.1131	.1112	.1093	.1075	.1056	.1038	.1020	.1003	.0985
1.30	.0968	.0951	.0934	.0918	.0901	.0885	.0869	.0853	.0838	.0823
1.40	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681
1.50	$\boxed{.0668}$	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
1.60	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
1.70	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
1.80	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
1.90	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
2.00	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
2.10	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
2.20	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
2.30	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
2.40	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064
2.50	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
2.60	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
2.70	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
2.80	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
2.90	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
3.00	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010

z	Area
3.500	.00023263
4.000	.00003167
4.500	.00000340
5.000	.00000029

Source: Computed by J. W. Stegeman using SAS.

Page A-3 OTT 1998
An Introduction To
STATISTICAL methods
and Data Analysis

Properties of the Normal Distribution

- Symmetric, bell-shaped density function.
- 68% of area under the curve between $\mu \pm \sigma$.
- 95% of area under the curve between $\mu \pm 2\sigma$.
- 99.7% of area under the curve between $\mu \pm 3\sigma$.

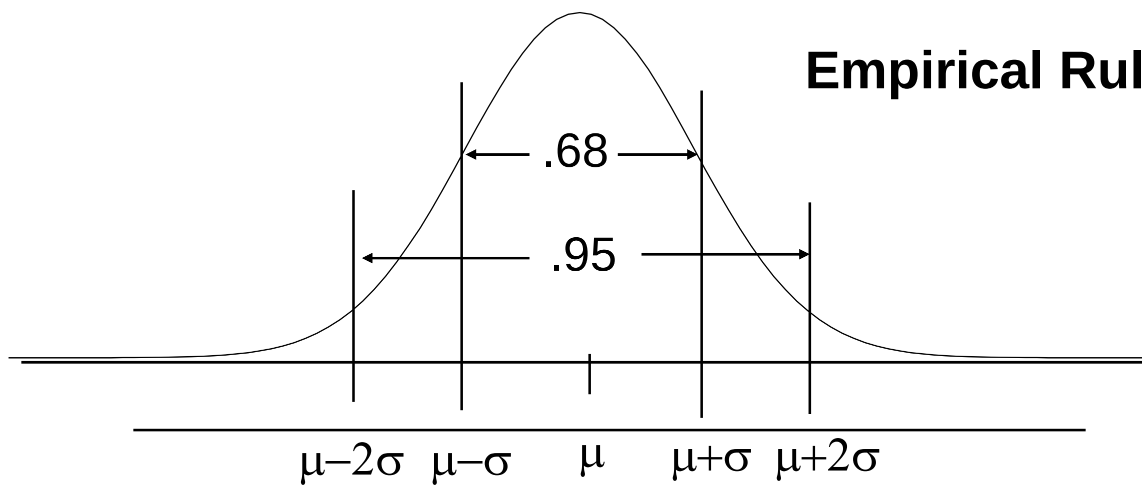
$$Y \sim N(\mu, \sigma)$$

$$Y - \mu \sim N(0, \sigma)$$

$$\frac{Y - \mu}{\sigma} \sim N(0,1)$$

Standard Normal Form

Empirical Rule



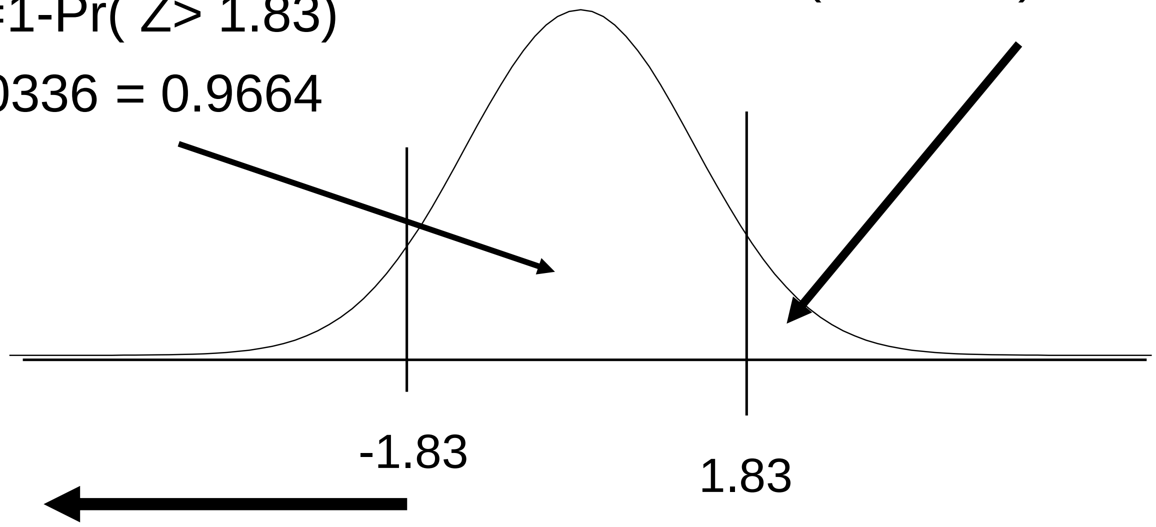
Probability Problems

Using symmetry and the fact that the area under the density curve is 1.

Pr(Z > 1.83) = 0.0336

Pr(Z < 1.83)=1-Pr(Z> 1.83)

=1-0.0336 = 0.9664



Pr(Z < -1.83) = Pr(Z> 1.83)

=0.0336

By Symmetry

Probability Problems

Cutting out the tails.

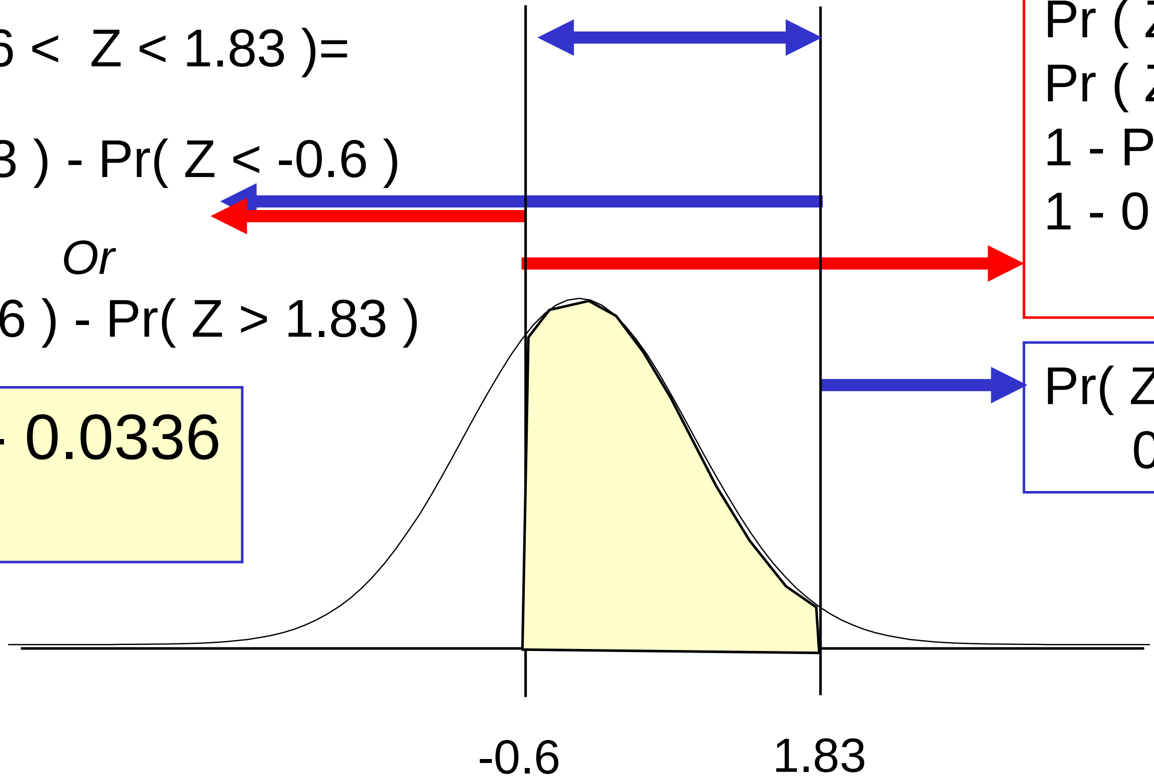
Pr(-0.6 < Z < 1.83)=

Pr(Z < 1.83) - Pr(Z < -0.6)

Or

Pr(Z > -0.6) - Pr(Z > 1.83)

= 0.7257 - 0.0336
= 0.6921



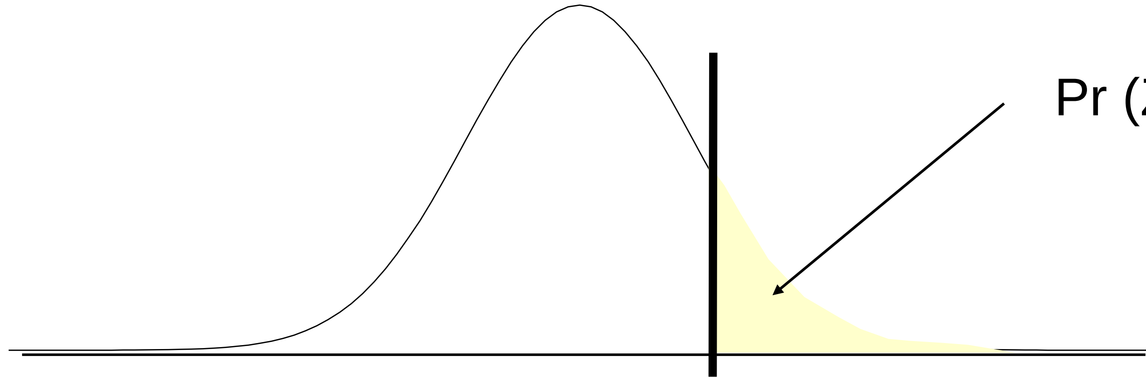
Pr (Z > -0.6) =
Pr (Z < +0.6) =
1 - Pr (Z > 0.6) =
1 - 0.2743 =
0.7257

Pr(Z > 1.83) =
0.0336

Given the Probability - What is Z_0 ?

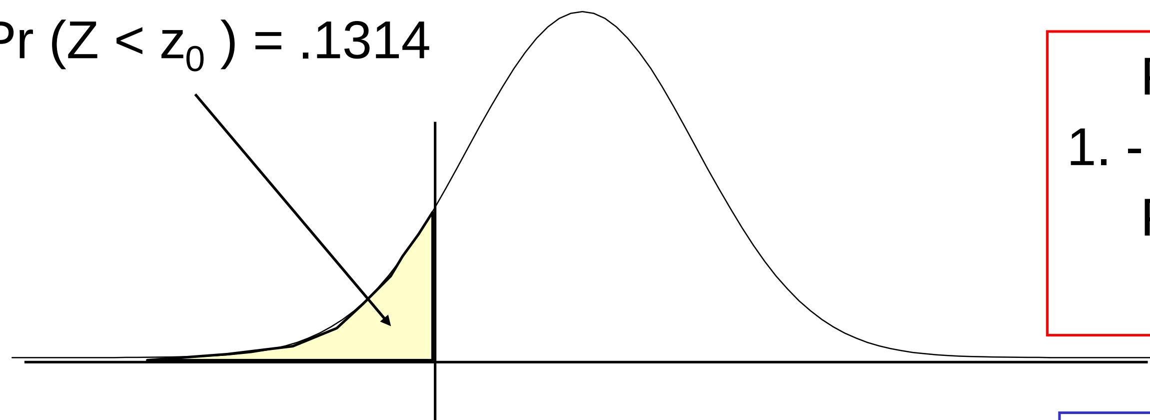
Working backwards.

$Pr\ (Z > z_0\ ) = .1314$



$z_0 = 1.12$

$Pr\ (Z < z_0\ ) = .1314$



z_0

$= -1.12$

$Pr\ (\ Z < z_0\ ) = 0.1314$

$1. - Pr\ (Z > z_0\ ) = 0.1314$

$Pr\ (\ Z > z_0\ ) = 1 - 0.1314$

$= 0.8686$

z_0 must be negative,
since $Pr\ (\ Z > 0.0\ ) = 0.5$

Converting to Standard Normal Form

$$Z = \frac{X - \mu}{\sigma}$$

Suppose we have a random variable (say weight), denoted by W, that has a normal distribution with mean 100 and standard deviation 10.

$W \sim N(100, 10)$

$Pr(W < 90) = Pr(Z < (90 - 100) / 10) = Pr(Z < -1.0)$

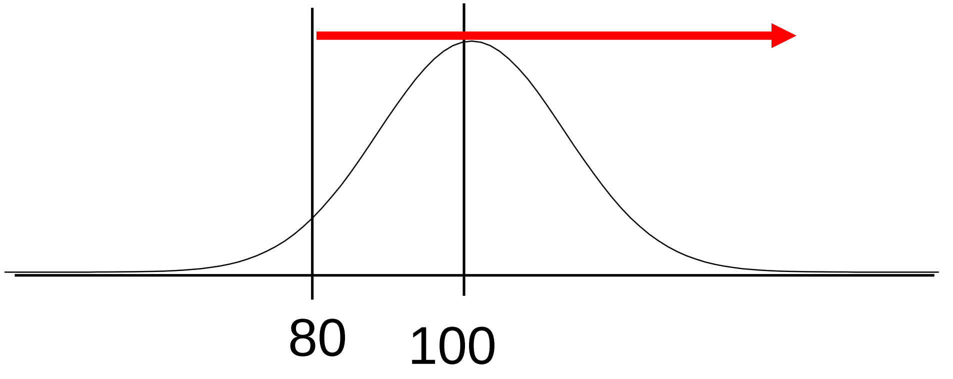
$= Pr(Z > 1.0) = 0.1587$

$Pr(W > 80) = Pr(Z > (80 - 100) / 10) = Pr(Z > -2.0)$

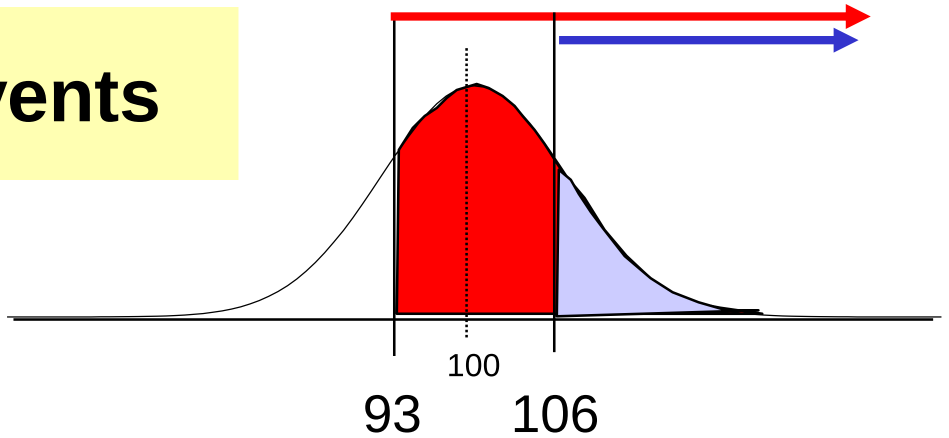
$= 1.0 - Pr(Z < -2.0)$

$= 1.0 - Pr(Z > 2.0)$

$= 1.0 - 0.0228 = 0.9772$



Decomposing Events



$W \sim N(100, 10)$

$Pr(93 < W < 106) = Pr(W > 93) - Pr(W > 106)$

$= Pr[Z > (93 - 100) / 10] - Pr[Z > (106 - 100) / 10]$

$= Pr[Z > -0.7] - Pr[Z > +0.6]$

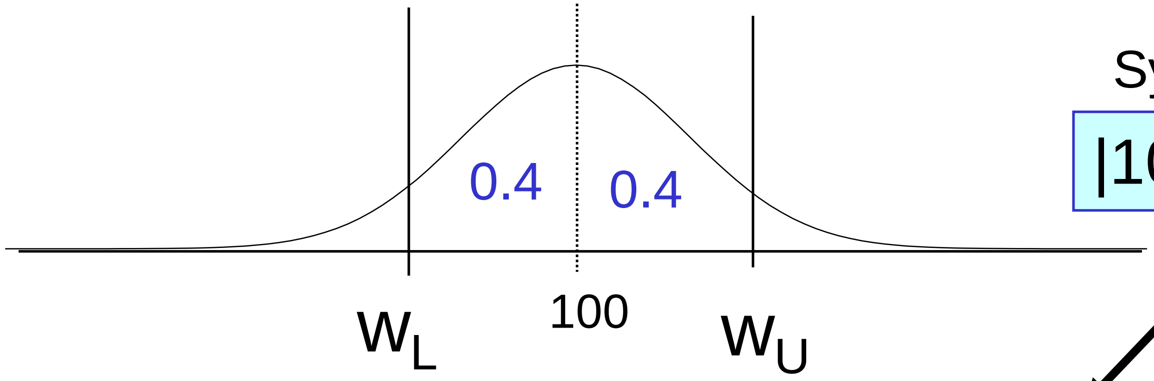
$= 1.0 - Pr[Z > +0.7] - Pr[Z > +0.6]$

$= 1.0 - 0.2420 - .2743 = 0.4837$

Finding Interval Endpoints

$W \sim N(100, 10)$

What are the two endpoints for a symmetric area centered on the mean that contains probability of 0.8 ?



Symmetry requirement

$$|100 - w_L| = |100 - w_U|$$

$$\mathbf{Pr(w_L < W < w_U) = 0.8}$$

$$Pr(100 < W < w_U) = Pr(w_L < W < 100) = 0.4$$

$$Pr(100 < W < w_U) = Pr(W > 100) - Pr(W > w_U) =$$
$$Pr(Z > 0) - Pr(Z > (w_U - 100)/10) = .5 - Pr(Z > (w_U - 100)/10) = 0.4$$

$$Pr(Z > (w_U - 100)/10) = 0.1 \Rightarrow (w_U - 100)/10 = z_{0.1} \approx 1.28$$

$$\boxed{w_U = 1.28 * 10 + 100 = 112.8}$$

$$\boxed{w_L = -1.28 * 10 + 100 = 87.2}$$

Probability Practice

Using Table 1 in Ott &Longnecker

Read probability in table using row (. 4) + column (.07) indicators.

Pr(Z < .47) = .6808

Pr(Z > .47) = 1-P(Z<.47)= 1-.6808=.3192

Pr (Z < -.47) = .3192

Pr (Z > -.47) = 1.0 - Pr (Z < -.47) = 1.0 - .3192 = .6808

Pr(.21 < Z < 1.56) = Pr (Z <1.56) - Pr (Z < .21)

.9406 - 0.5832 = .3574

Pr(-.21 < Z < 1.23) = Pr (Z < 1.23) - Pr (Z < -.21)

.8907 - .4168 = .4739

Finding Critical Values from the Table

Find probability in the Table, then read off row and column values.

$Pr(Z > z_{.2912}) = 0.2912$ $\quad z_{.2912} = \boxed{0.55}$

$Pr(Z > z_{.05}) = 0.05$ $\quad z_{.05} = \boxed{1.645}$

$Pr(Z > z_{.025}) = 0.025$ $\quad z_{.025} = \boxed{1.96}$

$Pr(Z > z_{.01}) = 0.01$ $\quad z_{.01} = \boxed{2.326}$