基本統計觀念

Some Basic Probability Concepts

Random Variable

random variable:

A variable whose value is unknown until it is observed. The value of a random variable results from an experiment.

The term random variable implies the existence of some known or unknown probability distribution defined over the set of all possible values of that variable.

In contrast, an arbitrary variable does not have a probability distribution associated with its values.

Controlled experiment values
of explanatory variables are chosen
with great care in accordance with
an appropriate experimental design.

Uncontrolled experiment values
of explanatory variables consist of
nonexperimental observations over
which the analyst has no control.

Discrete Random Variable

discrete random variable:
A discrete random variable can take only a finite number of values, that can be counted by using the positive integers.

Example: Prize money from the following lottery is a discrete random variable:

first prize: \$1,000
second prize: \$50
third prize: \$5.75

since it has only four (a finite number) (count: 1,2,3,4) of possible outcomes:

\$0.00; \$5.75; \$50.00; \$1,000.00

Continuous Random Variable

continuous random variable:
A continuous random variable can take any real value (not just whole numbers) in at least one interval on the real line.

Examples:

- Gross national product (GNP)
- money supply
- interest rates
- price of eggs
- household income
- expenditure on clothing

Dummy Variable

A discrete random variable that is restricted to two possible values (usually 0 and 1) is called a **dummy variable** (also, binary or indicator variable).

Dummy variables account for qualitative differences:

gender (0=male, 1=female),
race (0=white, 1=nonwhite),
citizenship (0=U.S., 1=not U.S.),
income class (0=poor, 1=rich).

A list of all of the possible values taken by a discrete random variable along with their chances of occurring is called a probability function or probability density function (pdf).

die	x	f(x)
one dot	1	1/6
two dots	2	1/6
three dots	3	1/6
four dots	4	1/6
five dots	5	1/6
six dots	6	1/6

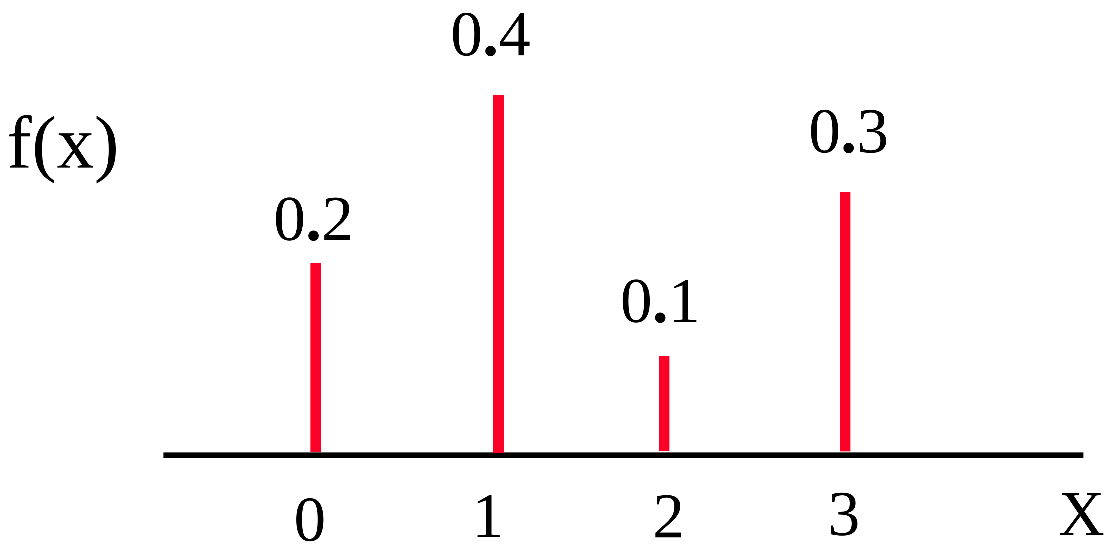
A discrete random variable X
has pdf, f(x), which is the **probability**
that X takes on the value x.

$$f(x) = P(X=x)$$

Therefore, $0 \leq f(x) \leq 1$

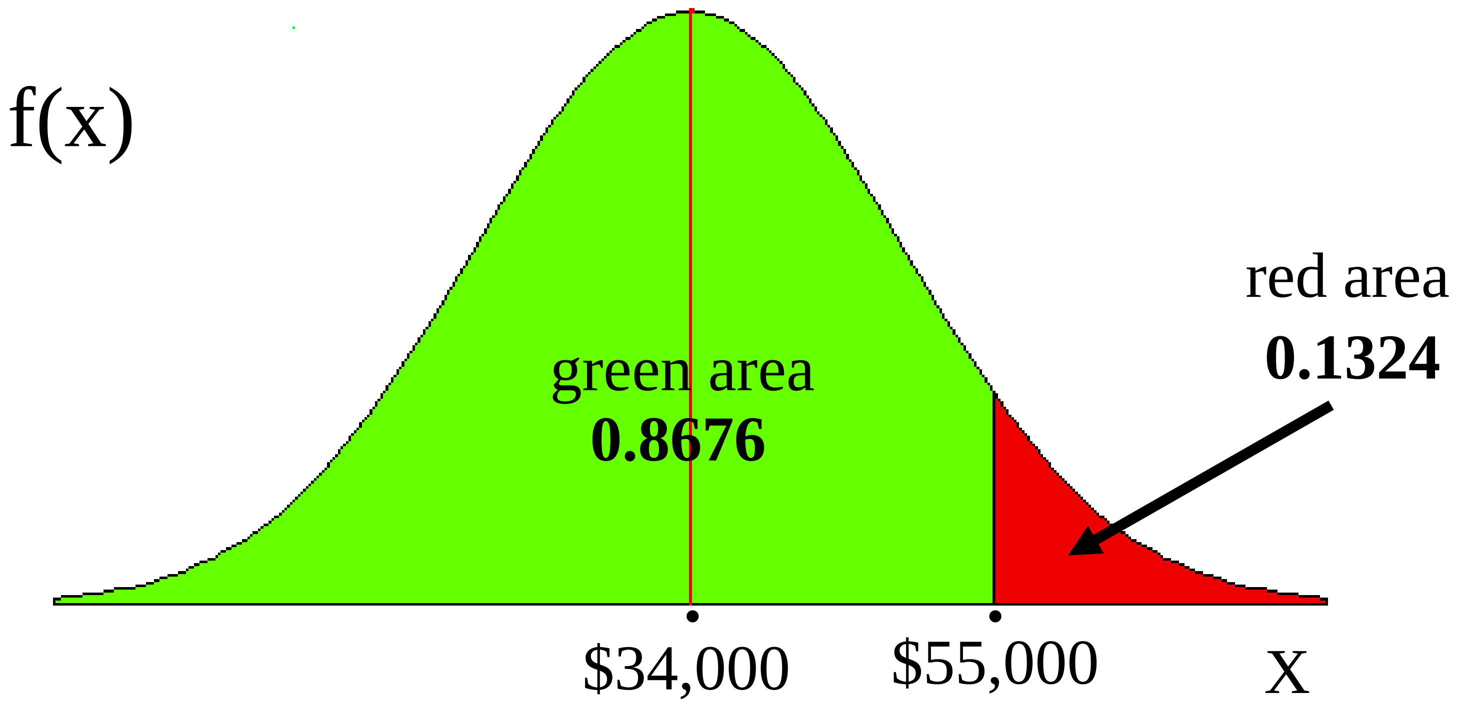
If X takes on the n values: $x_1, x_2, \ldots, x_n$,
then $f(x_1) + f(x_2) + \ldots + f(x_n) = 1$.

Probability, f(x), for a discrete random variable, X, can be represented by **height**:



number, X, on Dean's List of three roommates

A continuous random variable uses **area** under a curve rather than the height, f(x), to represent probability:



per capita income, X, in the United States

Since a continuous random variable has an **uncountably infinite** number of values, the probability of one occurring is **zero**.

$$P[\,X = a\,] \;=\; P[\,a \le X \le a\,] = 0$$

Probability is represented by **area**.

Height alone has no **area**.

An interval for X is needed to get an **area** under the curve.

The area under a curve is the integral of the equation that generates the curve:

$$P[\,a \le X \le b\,] = \int_a^b f(x)\,dx$$

For continuous random variables it is the integral of f(x), and not f(x) itself, which defines the area and, therefore, the probability.

Rules of Summation

Rule 1: $$\sum_{i=1}^{n} x_i = x_1 + x_2 + \ldots + x_n$$

Rule 2: $$\sum_{i=1}^{n} a x_i = a \sum_{i=1}^{n} x_i$$

Rule 3: $$\sum_{i=1}^{n} (x_i + y_i) = \sum_{i=1}^{n} x_i + \sum_{i=1}^{n} y_i$$

Note that summation is a linear operator which means it operates term by term.

Rules of Summation (continued)

Rule 4: $$\sum_{i=1}^{n} (ax_i + by_i) = a\sum_{i=1}^{n} x_i + b\sum_{i=1}^{n} y_i$$

Rule 5: $$\bar{x} = \frac{1}{n}\sum_{i=1}^{n} x_i = \frac{x_1 + x_2 + \ldots + x_n}{n}$$

The definition of $\bar{x}$ as given in Rule 5 implies the following important fact:

$$\sum_{i=1}^{n} (x_i - \bar{x}) = 0$$

Rules of Summation (continued)

Rule 6: $$\sum_{i=1}^{n} f(x_i) = f(x_1) + f(x_2) + \ldots + f(x_n)$$

Notation: $$\sum_{x} f(x_i) = \sum_{i} f(x_i) = \sum_{i=1}^{n} f(x_i)$$

Rule 7: $$\sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i,y_j) = \sum_{i=1}^{n} [\, f(x_i,y_1) + f(x_i,y_2) + \ldots + f(x_i,y_m)]$$

The order of summation does not matter **:**

$$\sum_{i=1}^{n} \sum_{j=1}^{m} f(x_i,y_j) = \sum_{j=1}^{m} \sum_{i=1}^{n} f(x_i,y_j)$$

The Mean of a Random Variable

The **mean** or arithmetic average of a random variable is its mathematical expectation or expected value, EX.

Expected Value

There are two entirely different, but mathematically equivalent, ways of determining the expected value:

1. Empirically:
   The **expected value** of a random variable, X, is the average value of the random variable in an infinite number of repetitions of the experiment.

In other words, draw an infinite number of samples, and average the values of X that you get.

Expected Value

2. Analytically:

The **expected value** of a discrete random variable, X, is determined by weighting all the possible values of X by the corresponding probability density function values, f(x), and summing them up.

In other words:

$$E[X] = x_1 f(x_1) + x_2 f(x_2) + \ldots + x_n f(x_n)$$

Empirical vs. Analytical

As sample size goes to infinity, the empirical and analytical methods will produce the same value.

In the empirical case when the sample goes to infinity the values of X occur with a frequency equal to the corresponding f(x) in the analytical expression.

Empirical (sample) mean:

$$\bar{x} = (1/n) \sum_{i=1}^{n} x_i$$

where **n** is the number of sample observations.

Analytical mean:

$$E[X] = \sum_{i=1}^{n} x_i \, f(x_i)$$

where **n** is the number of possible values of x_i.

Notice how the meaning of n changes.

The expected value of X:

$$E\,X = \sum_{i=1}^{n} x_i\, f(x_i)$$

The expected value of X-squared:

$$E\,X^2 = \sum_{i=1}^{n} x_i^2\, f(x_i)$$

It is important to notice that $f(x_i)$ does not change!

The expected value of X-cubed:

$$E\,X^3 = \sum_{i=1}^{n} x_i^3\, f(x_i)$$

$$EX = 0\,(.1) + 1\,(.3) + 2\,(.3) + 3\,(.2) + 4\,(.1)$$
$$= 1.9$$

$$EX^2 = 0^2(.1) + 1^2(.3) + 2^2(.3) + 3^2(.2) + 4^2(.1)$$
$$= 0 + .3 + 1.2 + 1.8 + 1.6$$
$$= 4.9$$

$$EX^3 = 0^3(.1) + 1^3(.3) + 2^3(.3) + 3^3(.2) + 4^3(.1)$$
$$= 0 + .3 + 2.4 + 5.4 + 6.4$$
$$= 14.5$$

$$E[g(X)] = \sum_{i=1}^{n} g(x_i)\ f(x_i)$$

$$g(X) = g_1(X) + g_2(X)$$

$$E[g(X)] = \sum_{i=1}^{n} [g_1(x_i) + g_2(x_i)]\ f(x_i)$$

$$E[g(X)] = \sum_{i=1}^{n} g_1(x_i)\ f(x_i) + \sum_{i=1}^{n} g_2(x_i)\ f(x_i)$$

$$E[g(X)] = E[g_1(X)] + E[g_2(X)]$$

Adding and Subtracting Random Variables

$$E(X+Y) = E(X) + E(Y)$$

$$E(X-Y) = E(X) - E(Y)$$

Adding a **constant** to a variable will add a constant to its expected value:

$$E(X+a) = E(X) + a$$

Multiplying by constant will multiply its expected value by that constant:

$$E(bX) = b\,E(X)$$

Variance

var(X) = average squared deviations around the mean of X.

var(X) = expected value of the squared deviations around the expected value of X.

$$\mathrm{var}(X) = E\,[(X - EX)^2]$$

$$\boxed{\mathrm{var}(X) = E\,[(X - EX)^2]}$$

$$\begin{aligned}
\mathrm{var}(X) &= E\,[(X - EX)^2] \\
&= E\,[X^2 - 2XEX + (EX)^2] \\
&= E(X^2) - 2\,EX\,EX + E\,(EX)^2 \\
&= E(X^2) - 2\,(EX)^2 + (EX)^2 \\
&= E(X^2) - (EX)^2
\end{aligned}$$

$$\boxed{\mathrm{var}(X) = E(X^2) - (EX)^2}$$

variance of a discrete random variable, X:

$$\mathrm{var}(X) = \sum_{i=1}^{n} (x_i - EX)^2 f(x_i)$$

standard deviation is square root of variance

calculate the <u>variance</u> for a discrete random variable, X:

x_i	$f(x_i)$	$(x_i - EX)$	$(x_i - EX)^2 f(x_i)$
2	.1	2 - 4.3 = -2.3	5.29 (.1) = .529
3	.3	3 - 4.3 = -1.3	1.69 (.3) = .507
4	.1	4 - 4.3 = - .3	.09 (.1) = .009
5	.2	5 - 4.3 = .7	.49 (.2) = .098
6	.3	6 - 4.3 = 1.7	2.89 (.3) = .867

$$\sum_{i=1}^{n} x_i f(x_i) = .2 + .9 + .4 + 1.0 + 1.8 = 4.3$$

$$\sum_{i=1}^{n} (x_i - EX)^2 f(x_i) = .529 + .507 + .009 + .098 + .867$$
$$= 2.01$$

$$Z = a + cX$$

$$\begin{aligned}\mathrm{var}(Z) &= \mathrm{var}(a + cX) \\ &= E\,[(a+cX) - E(a+cX)]^2 \\ &= c^2\ \mathrm{var}(X)\end{aligned}$$

$$\boxed{\mathrm{var}(a + cX) = c^2 \mathrm{var}(X)}$$

Covariance

The **covariance** between two random variables, X and Y, measures the linear association between them.

$$\mathrm{cov}(X,Y) = E[(X - EX)(Y-EY)]$$

Note that variance is a special case of covariance.

$$\mathrm{cov}(X,X) = \mathrm{var}(X) = E[(X - EX)^2]$$

$$\boxed{\mathrm{cov}(X,Y) = E\,[(X - EX)(Y-EY)]}$$

$$\begin{aligned}
\mathrm{cov}(X,Y) &= E\,[(X - EX)(Y-EY)] \\
&= E\,[XY - X\,EY - Y\,EX + EX\,EY] \\
&= E(XY) - EX\,EY - EY\,EX + EX\,EY \\
&= E(XY) - 2\,EX\,EY + EX\,EY \\
&= E(XY) - EX\,EY
\end{aligned}$$

$$\boxed{\mathrm{cov}(X,Y) = E(XY) - EX\,EY}$$

	Y = 1	Y = 2	
X = 0	.45	.15	.60
X = 1	.05	.35	.40
	.50	.50	

EX=0(.60)+1(.40)=.40

covariance

EY=1(.50)+2(.50)=1.50

cov(X,Y) = E(XY) - EX EY
= .75 - (.40)(1.50)
= .75 - .60
= .15

EX EY = (.40)(1.50) = .60

E(XY) = (0)(1)(.45)+(0)(2)(.15)+(1)(1)(.05)+(1)(2)(.35)=.75

Correlation

The **correlation** between two random variables X and Y is their covariance divided by the square roots of their respective variances.

$$\rho(X,Y) = \frac{\operatorname{cov}(X,Y)}{\sqrt{\operatorname{var}(X)}\sqrt{\operatorname{var}(Y)}}$$

Correlation is a pure number falling **between -1 and 1**.

	Y = 1	Y = 2	
X = 0	.45	.15	.60
X = 1	.05	.35	.40
	.50	.50	

$EX = .40$

$EX^2 = 0^2(.60) + 1^2(.40) = .40$

$var(X) = E(X^2) - (EX)^2$
$= .40 - (.40)^2$
$= .24$

$cov(X,Y) = .15$

correlation

$EY = 1.50$

$EY^2 = 1^2(.50) + 2^2(.50)$
$= .50 + 2.0$
$= 2.50$

$var(Y) = E(Y^2) - (EY)^2$
$= 2.50 - (1.50)^2$
$= .25$

$$\rho(X,Y) = \frac{cov(X,Y)}{\sqrt{var(X)\ var(Y)}}$$

$$\boxed{\rho(X,Y) = .61}$$

Zero Covariance & Correlation

Independent random variables have zero covariance and, therefore, zero correlation.

The converse is not true.

Since expectation is a linear operator, it can be applied term by term.

The expected value of the weighted sum of random variables is the sum of the expectations of the individual terms.

$$E[c_1X + c_2Y] = c_1EX + c_2EY$$

In general, for random variables $X_1, \ldots, X_n$:

$$E[c_1X_1 + \ldots + c_nX_n] = c_1EX_1 + \ldots + c_nEX_n$$

The **variance of a weighted sum** of random variables is the sum of the variances, each times the square of the weight, plus twice the covariances of all the random variables times the products of their weights.

Weighted **sum** of random variables:

$$\mathrm{var}(c_1X + c_2Y) = c_1^2\,\mathrm{var}(X) + c_2^2\,\mathrm{var}(Y) + 2c_1c_2\,\mathrm{cov}(X,Y)$$

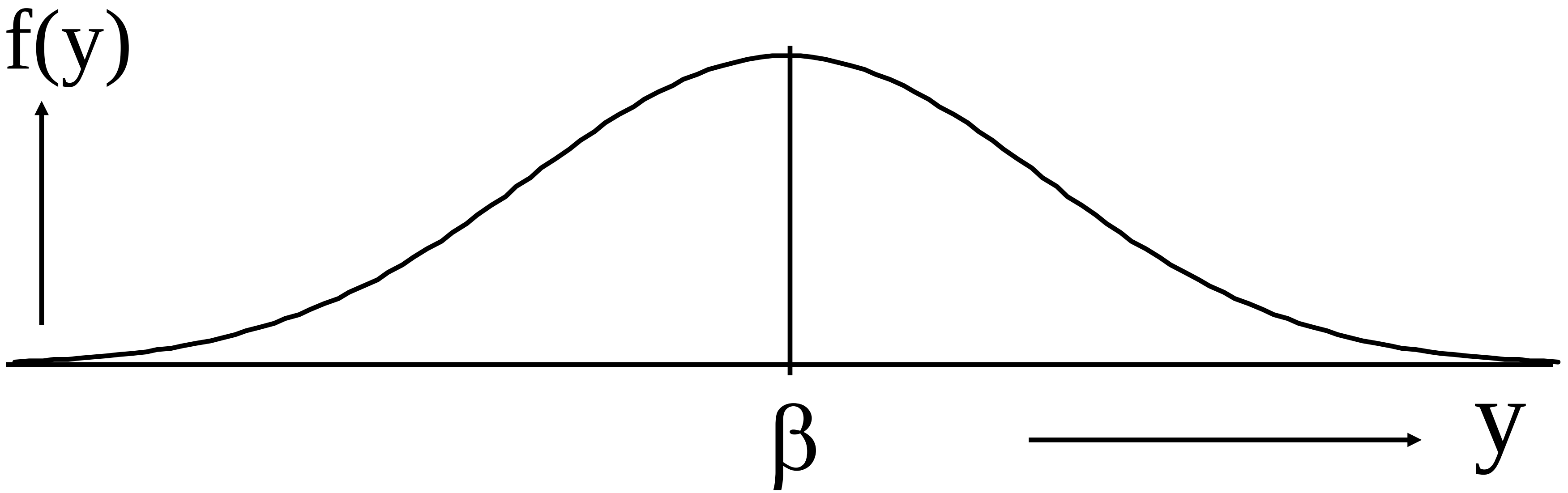
Weighted **difference** of random variables:

$$\mathrm{var}(c_1X - c_2Y) = c_1^2\,\mathrm{var}(X) + c_2^2\,\mathrm{var}(Y) - 2c_1c_2\,\mathrm{cov}(X,Y)$$

The Normal Distribution

$$Y \sim N(\beta, \sigma^2)$$

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y-\beta)^2}{2\sigma^2}\right]$$



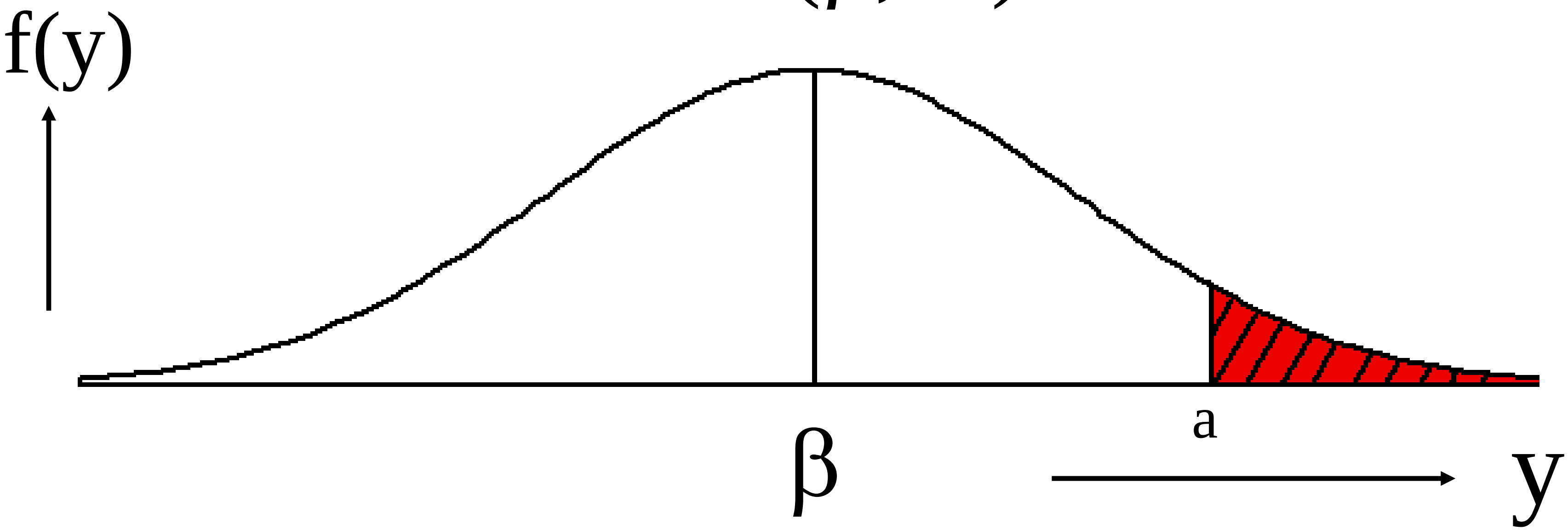
The Standardized Normal

$$Z = (y - \beta)/\sigma$$

$$Z \sim N(0,1)$$

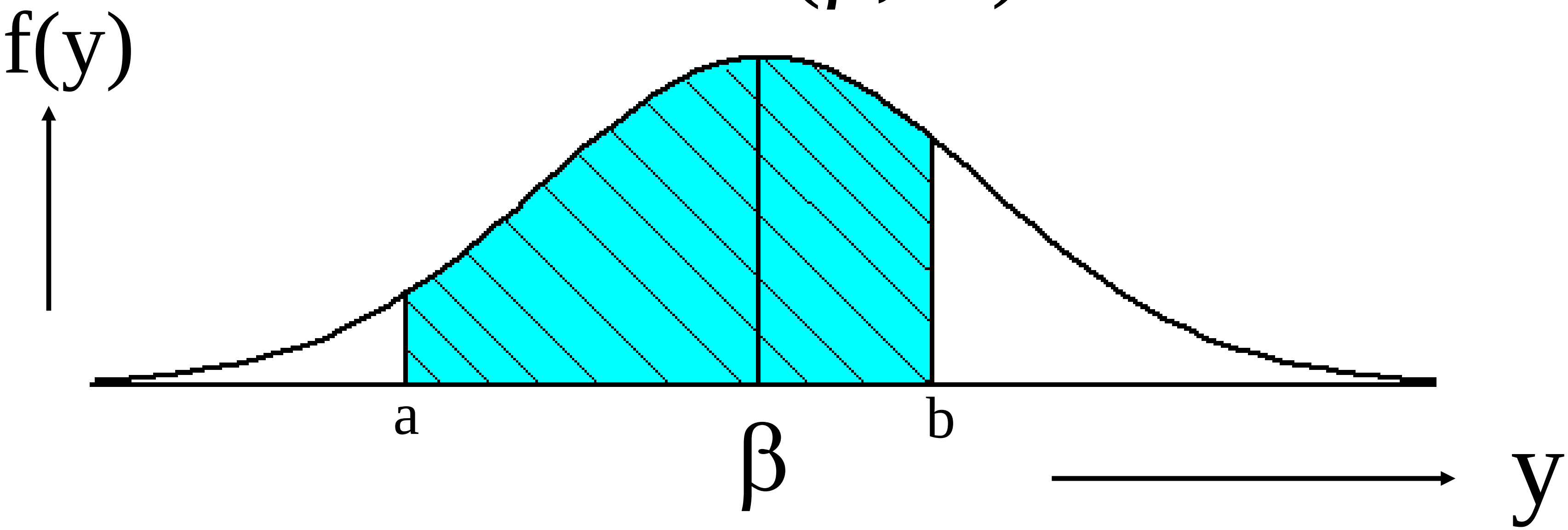
$$f(z) = \frac{1}{\sqrt{2\pi}} \exp\left[\frac{-z^2}{2}\right]$$

$$Y \sim N(\beta, \sigma^2)$$



$$\mathbf{P}\left[\,\mathbf{Y} \geq a\,\right] = \mathbf{P}\left[\frac{Y - \beta}{\sigma} \geq \frac{a - \beta}{\sigma}\right] = \mathbf{P}\left[\mathbf{Z} \geq \frac{a - \beta}{\sigma}\right]$$

$$Y \sim N(\beta,\sigma^2)$$



$$P\left[a \le Y \le b\right] = P\left[\frac{a-\beta}{\sigma} \le \frac{Y-\beta}{\sigma} \le \frac{b-\beta}{\sigma}\right]$$

$$= P\left[\frac{a-\beta}{\sigma} \le Z \le \frac{b-\beta}{\sigma}\right]$$

Linear combinations of jointly normally distributed random variables are themselves normally distributed.

$$Y_1 \sim N(\beta_1,\sigma_1^2),\ Y_2 \sim N(\beta_2,\sigma_2^2),\ \ldots,\ Y_n \sim N(\beta_n,\sigma_n^2)$$

$$W = c_1 Y_1 + c_2 Y_2 + \ldots + c_n Y_n$$

$$W \sim N[\ E(W), \mathrm{var}(W)\ ]$$

Chi-Square

If $Z_1, Z_2, \ldots, Z_m$ denote m independent N(0,1) random variables, and

$V = Z_1^2 + Z_2^2 + \ldots + Z_m^2$, then $V \sim \chi^2_{(m)}$

V is **chi-square** with m degrees of freedom.

mean: $E[V] = E[\chi^2_{(m)}] = m$

variance: $\mathrm{var}[V] = \mathrm{var}[\chi^2_{(m)}] = 2m$

Student - t

If $Z \sim N(0,1)$ and $V \sim \chi^2_{(m)}$ and if Z and V are independent then,

$$t = \frac{Z}{\sqrt{V/m}} \sim t_{(m)}$$

t is **student-t** with m degrees of freedom.

mean: $E[t] = E[t_{(m)}] = 0$ symmetric about zero

variance: $\mathrm{var}[t] = \mathrm{var}[t_{(m)}] = m/(m-2)$

F Statistic

If $V_1 \sim \chi^2_{(m_1)}$ and $V_2 \sim \chi^2_{(m_2)}$ and if V_1 and V_2 are independent, then

$$F = \frac{V_1/m_1}{V_2/m_2} \sim F_{(m_1,m_2)}$$

F is an **F statistic** with m_1 numerator degrees of freedom and m_2 denominator degrees of freedom.