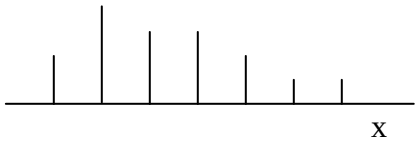
8. Continuous Distributions - the Normal (Gaussian)

Readings: **Rosner: 5.1-6**
SAS Functions: RANNOR, PROBNORM, PROBIT
R: pnorm, qnorm, dnorm, rnorm

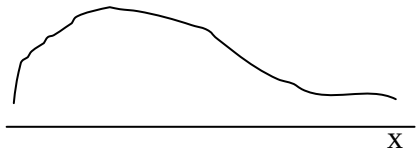
A) Review of Continuous Distributions
B) The Normal Distribution
C) Standard Normal Distribution and probabilities

A) General Properties of Continuous Distributions

For discrete quantities, we use a probability mass function for the probability distribution of X: Probability distribution of X is $p(x) = P(X = x)$



For continuous quantities, we use a probability density function: pdf or p(x) or f(x)

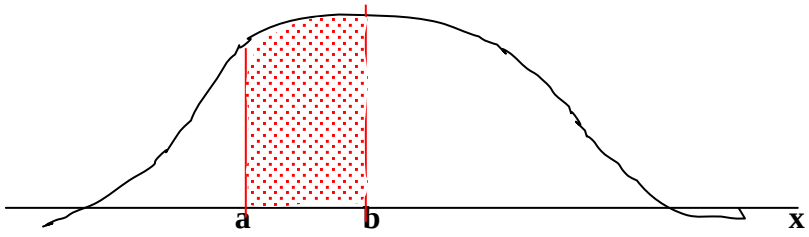


Continuous distributions have the following properties:

1) $P(a \le X \le b)$ = area under curve from a to b is equal to the probability that the random variable X falls between a and b

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

$$P(a \le X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a)$$



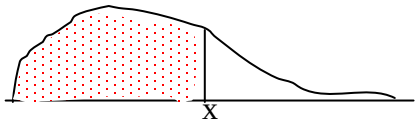
2) Probability of individual values is 0: P(X=a) = 0. Hence, no distinction made between P(X < a) and P (X ≤ a).

3) Total area under the curve over the entire range of possible values for the random variable is 1. Area under f(x) =

$$\int_{-\infty}^{\infty} f(x)dx = 1$$

4) Cumulative distribution function (cdf):

$$F(x) = P(X \le x) = P(X < x) = \int_{-\infty}^{x} f(u)du$$



e.g. Suppose diastolic blood pressure is symmetrically distributed around 80 mmHg.

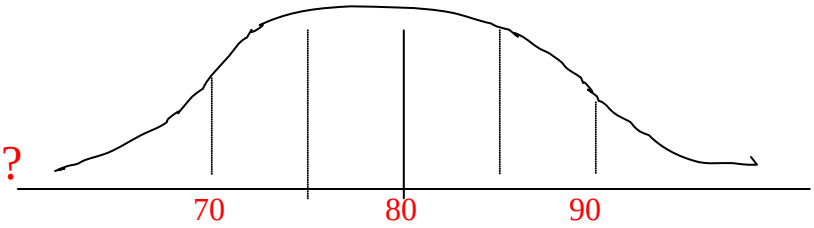
Which probabilities are largest or smallest?

P(DBP ≤ 70)?

P(DBP > 90)?

P(DBP >95)?

P(DBP < 70 or DBP > 90)?



5) Continuous random variables have means and variances (and standard deviations) that describe their location (center) and spread

Recall - **Expected Value:** Average value of the random variable

$$E[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_{\text{all } x} xf(x)dx = \mu$$

Recall - **Variance:** Average squared distance of each value of the random variable from its expected value:

$$V[X] = \int_{-\infty}^{\infty} (x-\mu)^2 f(x)dx = \int_{\text{all } x} (x-\mu)^2 f(x)dx = \sigma^2$$

Standard Deviation: $s.d.(X) = \sqrt{V[X]} = \sigma$

6) If the distribution is roughly normal, the 68% and 95% rules that we saw before roughly hold.

e.g. Suppose DBP has mean 80 mmHg and s.d. 5 mmHg. Roughly what are the probabilities in the previous example?

B) The Normal Distribution (or Gaussian: K. F. Gauss, 1777-1855; Figure 5.4 in Rosner)

The normal distribution is the most important continuous distribution in statistics by far. This is because:

1) It has good theoretical properties (e.g. sums, differences, and averages of normal r.v. are normally distributed).

2) It is closely related to many other distributions (e.g. the binomial and Poisson distributions are approximately normal in certain circumstances).

3) Many naturally occurring variables roughly follow a normal distribution (e.g. body weights and heights, blood pressure, etc.).

4) Many variables that are right (positively) skewed can be transformed, usually by logs, to normal r.v. (e.g. plasma concentrations – cholesterol, blood lead, etc.).

The normal distribution is defined by its *probability density function* (pdf), which is given as:

$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2\sigma^2}(x-\mu)^2\right) \quad -\infty<x<\infty$$

$$\exp[\,]=e^{[\,]} \quad e\cong 2.7182 \quad \pi\cong 3.14159$$

for parametersμ, $\sigma^2>0$; sometimes the pdf is written as

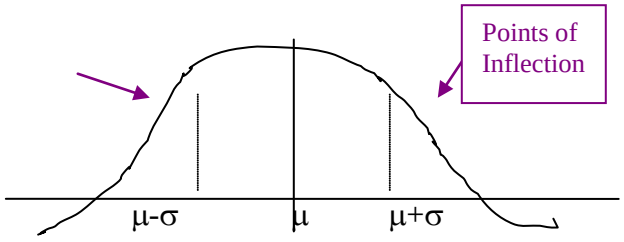
$$f(x)=\frac{1}{\sqrt{2\pi}\sigma}\exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right)$$

The parameters μ and σ^2 completely determine the normal distribution. These are (not coincidentally) the distribution's mean and variance. X ~N (μ, σ^2) means "X is distributed normal with mean μ and variance σ^2".

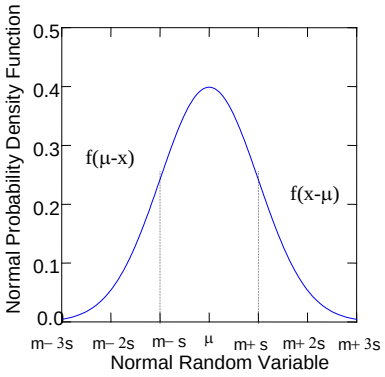
Mean: $\mathrm{E}[\mathrm{X}]=\int_{-\infty}^{\infty} x f(x)dx=\mu$

Variance: $\mathrm{V}[\mathrm{X}]=\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx=\sigma^2$

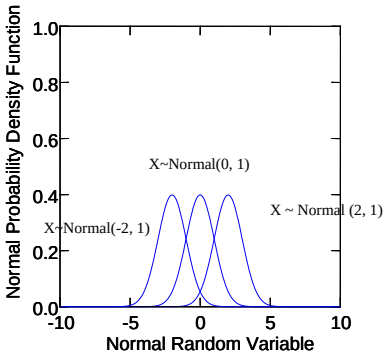
μ is the most frequently occurring value in the distribution and the curve is symmetric about μ: i.e. f(x-μ) = f(μ-x). Inflection points (where the curve (curvature) changes direction (sign)) occur at μ-σ, μ+σ. Note from the density formula that the height of *f(x)* is inversely proportional to the s.d., σ.



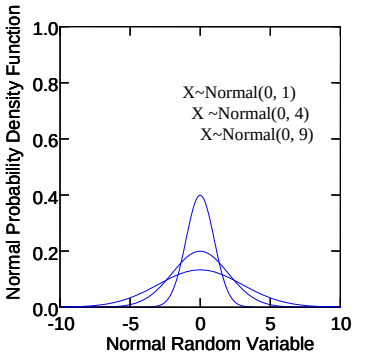
A generic normal distribution:



Normal distributions with the same variance but different means



Normal distributions with the same means but different variances:



Population Moments

The population moments can describe the location, variability, skewness, and kurtosis of a population, just as the corresponding sample moments describe a sample.

$E(X) = \mu$, $E(X^2)$ are population moments about zero
$E[(X-\mu)] = 0$, $E[(X-\mu)^2] = \sigma^2$, are population moments about μ (these are called *central* moments)

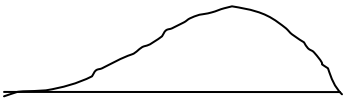
Skewness describes symmetry of distribution.

The 3rd central moment of the data, as with the 1st central moment, will balance out from left to right if the data are symmetric. With normally distributed data, we expect skewness to be 0.

$(E[(X-\mu)^3])$ standardized by $s^3 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^3$

If skewness is > 0: positive skew; skewed to the right
If skewness if < 0: negative skew; skewed to the left

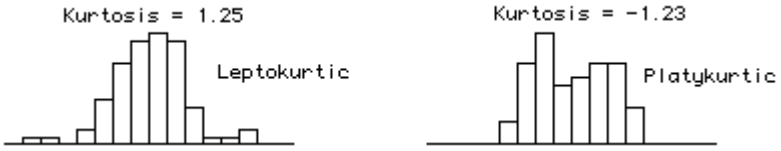

skewness > 0, positive, common


skewness < 0, negative

Kurtosis: 4th central moment standardized by $s^4 - 3$:

$s^4 = \frac{1}{n}\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^4$, 3 is kurtosis for a true normal distribution

kurtosis = 0, tails just like a normal distribution
kurtosis > 0 heavier tails than a normal distribution
kurtosis < 0 lighter tails than a normal distribution



> 0, heavier tails < 0, thinner tails

Contrary to popular belief, kurtosis has pretty much nothing to do with the peakedness of a distribution.

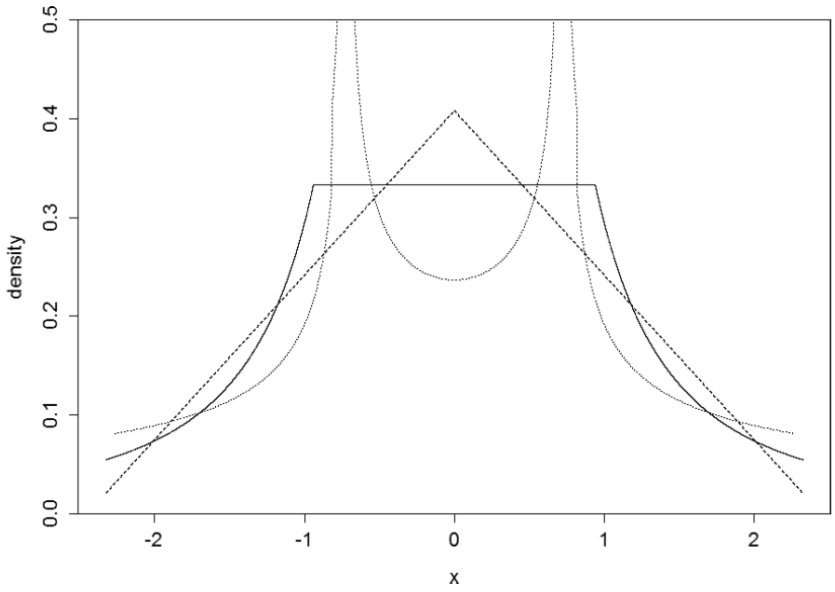


Figure 2.
Distributions with identical kurtosis = 2.4: solid = devil's tower, dashed = triangular, dotted = slip-dress.

Ref: Kurtosis as Peakedness, 1905-2014, *R.I.P.* Peter H. Westfall, *The American Statistician*, August, 68:191-195.

C) Standard Normal Distribution X ~ N (0, 1)

First we'll study the "standard normal" distribution:

$$N(0,1) \ (\mu = 0 \ , \ \sigma^2 = 1).$$

Next we'll see how to convert any N (μ, σ^2) r.v. to a standard normal r.v., then use the normal tables to find probabilities.

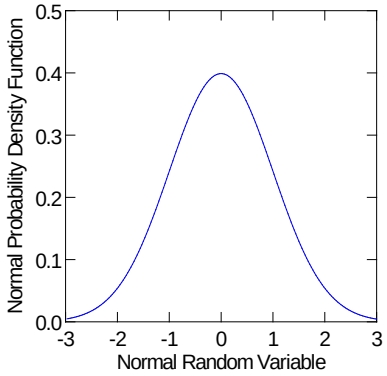
For a standard normal r.v.: $f(x) = \frac{1}{\sqrt{2\pi}} \exp\left(-\frac{1}{2} x^2 \right)$

And the distribution is symmetric about 0: f(x) = f(-x).

P(-1 < X < 1) =0.6827 = (about 68% of area lies between ± 1)

P(-1.96 < X < 1.96) = 0.95 (about 95% of area lies between ± 2)

P(-2.576 < X < 2.576) = 0.99 (about 99% of area lies between ± 2.5)



Thus absolute values greater than 3 are highly unlikely.

Transforming any normal distribution to the standard normal distribution is a simple case of *Translation* and *Re-scaling.*

Consider: X ~ Normal (μ , σ^2)

Recall that translation, X + c, results in a change in the mean, but not the variance or s.d.

$$X - \mu \sim \text{Normal}(\mu - \mu, \sigma^2) \sim \text{Normal}(0, \sigma^2)$$

Recall that re-scaling, cX, results in a change in the mean *and* in the variance (and s.d.)

$$Z = (X - \mu)/\sigma \quad \sim \quad \text{Normal}(0/\sigma, \sigma^2/\sigma^2) \sim \text{Normal}(0, 1)$$

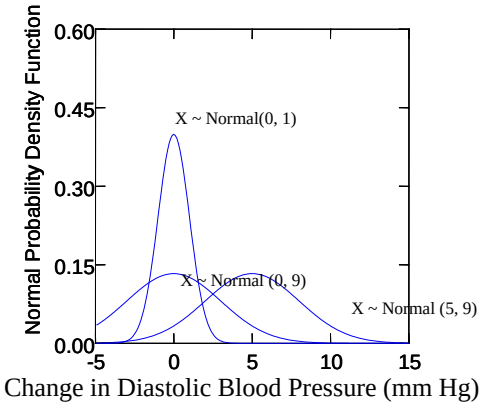
We call the standard normal distribution the Z distribution, $(X - \mu)/\sigma$ is a Z-score, and the act of translating and re-scaling a normal r.v. so that it has a Z distribution is called: "*Doing the Z-thing*!"

Note: Normality is preserved by translation and re-scaling.

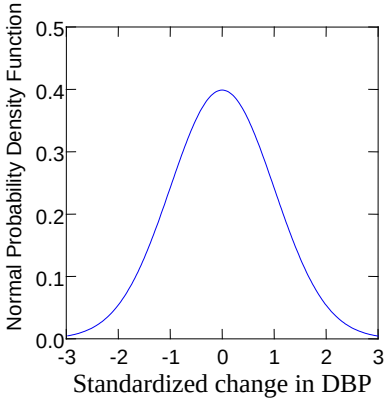
Example: Let X = the change (increase ☹? or decrease ☺?) in DBP after a class of BIOS 6611 ~ Normal (5, 9).

Any value in this distribution can be mapped to a value in the standard normal distribution.

So, first we translate by subtracting 5 mm Hg to get a Normal (0, 9) r.v., then we re-scale to get a Normal (0, 1) r.v., i.e. a standard normal r.v., or Z.



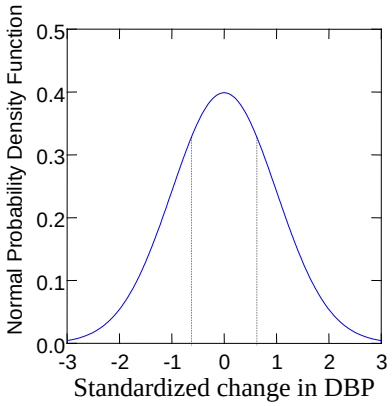
Using the Normal Probability Tables (Table 3 in Rosner; Jack's Tables):



How many s.d. units is 7 mm Hg away from the mean? 3 mm Hg?

What percent of the distribution is above 7 mm Hg? below 3 mm Hg?

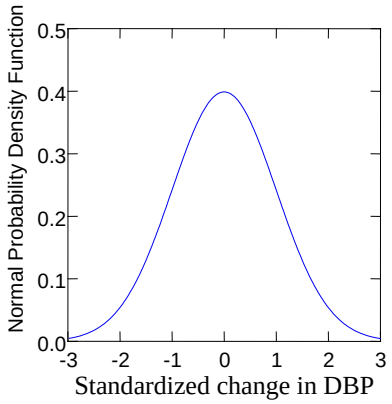
How do we convert 7 and 3 into values on the standardized scale?



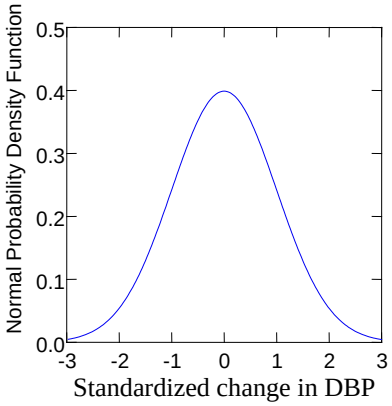
Shade in the areas of probability for the percent of the distribution above 7 mm Hg ...

for the percent below 3 mm Hg ...

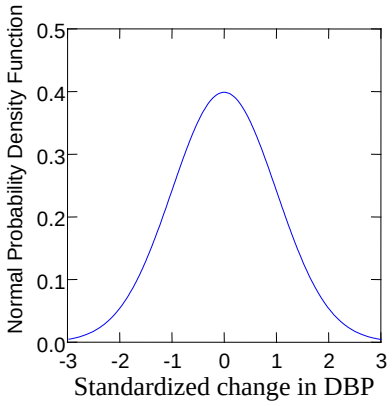
What percent of the distribution is between 3 and 7 mm Hg?



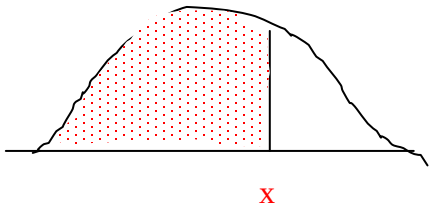
What percent is between 0 and 10 mm Hg?



Between what two values of DBP does the middle 50% of the distribution lie?



Cumulative distribution function: cdf for standard normal: $\Phi(x) = P(X \leq x)$



By symmetry: $\Phi(-x) = P(X \leq -x) = P(X \geq x) = 1 - P(X \leq x) = 1 - \Phi(x)$

Using Table 3 The Normal Distribution in Appendix of Rosner:

		Jack's Tables (z)
Column x:	values of x	z-score
Column A:	$\Phi(x) = P(X \leq x)$	Area below z-score
Column B:	$1 - \Phi(x) = P(X > x)$	Area above z-score
Column C:	$P(0 \leq X \leq x)$	----
Column D:	$P(-x \leq X \leq x)$	Area between +/-z=score

We can also use SAS to obtain probabilities and normal deviates (oxymoron?), or z values of the N (0, 1) distribution:

```
DATA probs;
    z = 2.12;
      prob3 = probnorm(z);
            /*gives P(Z < z) for Z~ N (0,1)*/
    p = .7;
      z3 = probit (p);
            /*gives z3 so that P(z < z3) = p*/
RUN;

PROC PRINT; RUN;
```

Obs	z	prob3	p	z3
1	2.12	0.98300	0.7	0.52440

Recall - **Percentiles:** values of x corresponding to cumulative the distribution function $\Phi(x)$ are known as the 100% $\Phi(x)$ percentiles

$\Phi(x) = 0.95$: Find the value of *x* such that 95% of observations are below it:

Look up 0.95 in column A of Table 3 and find associated *x*:

What percentile is $x = 1.2$?

Using SAS:

```
DATA probs;
    z = 1.2;
      prob3 = probnorm(z);
    p = .95;            /*gives P(Z < z) for Z~ N (0,1)*/
      z3 = probit (p);  /*gives z3 so that P(z < z3) = p*/
  RUN;


PROC PRINT; RUN;
```

Obs	z	prob3	p	z3
1	1.2	0.88493	0.95	1.64485

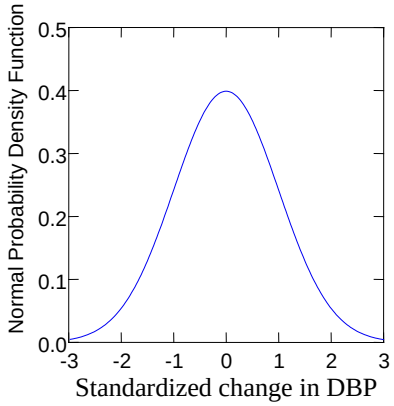
Useful percentiles of Z to know (and impress others with at social events!):

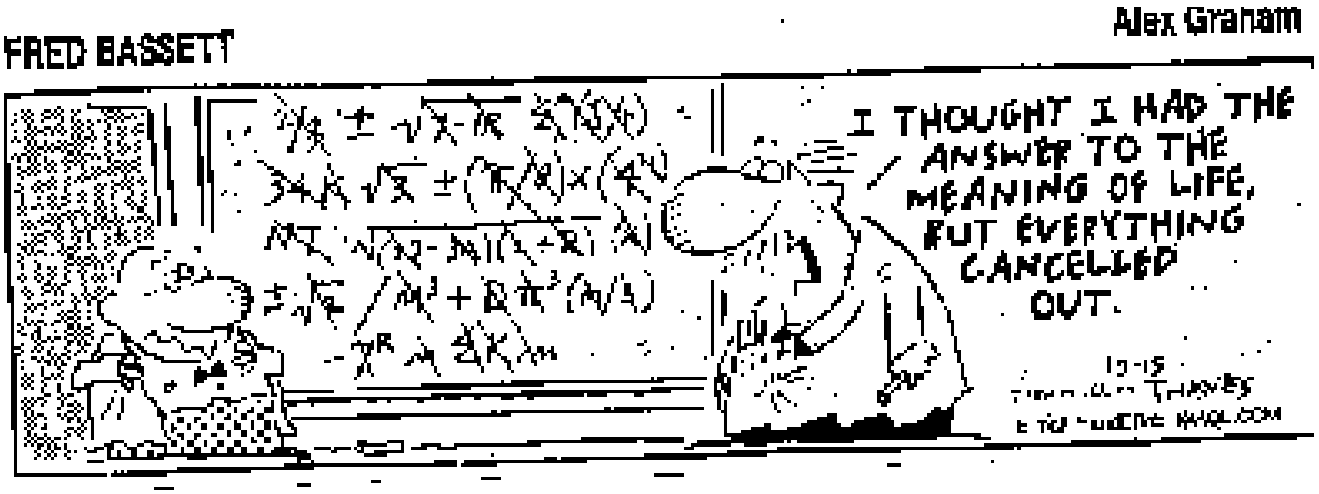
80th percentile =

90th percentile =

95th percentile =

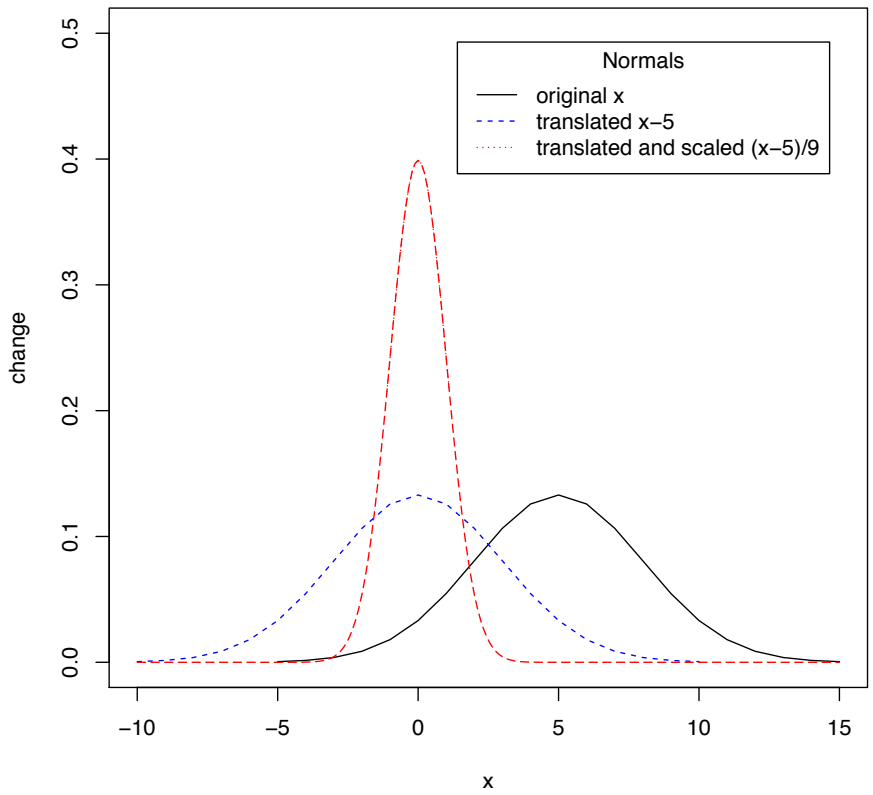
What are the 80th, 90th and 95th percentiles of the DBP change distribution?




FRED BASSETT
Alex Graham
I THOUGHT I HAD THE ANSWER TO THE MEANING OF LIFE, BUT EVERYTHING CANCELLED OUT.

Normal distributions, translated and scaled.

```
# unit 8
x <- c(-5:15)
x2 <- x - 5
x3 <- x2/9
x4 <- (-90:135)/9
change <- dnorm(x, mean = 5, sd = 3)
plot(x, change, type = "l", xlim = c(-10, 15), ylim = c(0, 0.5), lty = 1)
lines(x2, dnorm(x2, mean = 0, sd = 3), col = "blue", lty = 2)
lines(x3, dnorm(x3, mean = 0, sd = 1), col = "red", lty = 3)
lines(x4, dnorm(x4, mean = 0, sd = 1), col = "red", lty = 5)
legend("topright", inset = 0.05, title = "Normals", legend = c("original x",
    "translated x-5", "translated and scaled (x-5)/9"), col = c("black", "blue",
    "red"), lty = c(1, 2, 3))
```



Using *pnorm()* is similar to using SAS `PROBNORM`.

```
pnorm(2.12)

## [1] 0.983

qnorm(0.7)

## [1] 0.5244

z <- 1.2
pnorm(z)

## [1] 0.8849

p <- 0.95
qnorm(p)

## [1] 1.645
```

Appendix: Code

```
# unit 8
x<-c(-5:15)
x2<-x-5
x3<-x2/9
x4<-(-90:135)/9
change<-dnorm(x,mean=5,sd=3)
plot(x,change,type="l",xlim=c(-10,15),ylim=c(0,0.5),lty=1)
lines(x2,dnorm(x2,mean=0,sd=3),col="blue",lty=2)
lines(x3,dnorm(x3,mean=0,sd=1),col="red",lty=3)
lines(x4,dnorm(x4,mean=0,sd=1),col="red",lty=5)
legend("topright",inset=0.05,title="Normals",legend=c("original x","translated x-5
    ","translated and scaled (x-5)/9"),col=c("black","blue","red"),lty=c(1,2,3))
pnorm(2.12)
qnorm(.7)

z <- 1.2
pnorm(z)
p <- 0.95
qnorm(p)
```