Section 1.3 - Normal Distributions

Statistics 104

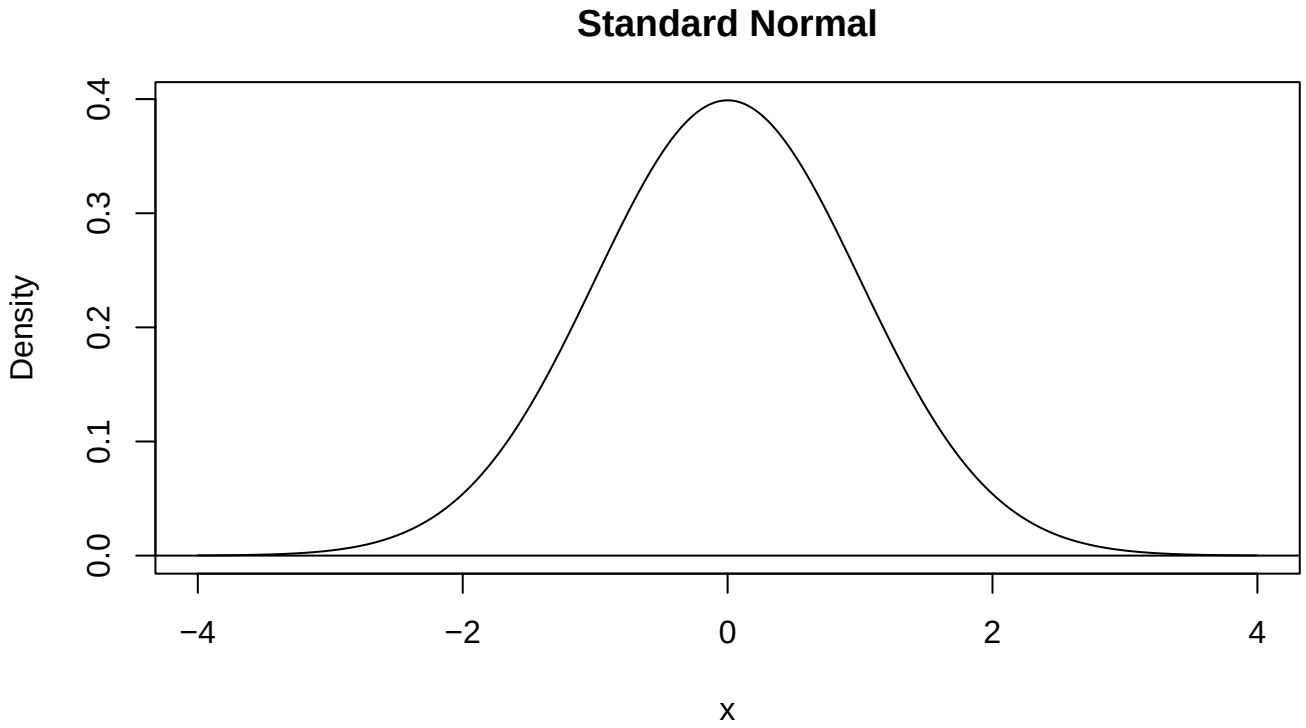
Autumn 2004

Normal Distributions

The normal distribution is almost surely the most common distribution used in probability and statistics. It is also referred to as the Gaussian distribution, as Gauss was an early promoter of its use (though not the first, which was probably De Moivre). It is also what most people mean when they talk about bell curve. It is used to describe observed data, measurement errors, an approximation distribution (Central Limit Theorem).

The density is defined by two parameters, the mean μ, and standard deviation σ. (The mean and standard deviation of a probability distribution will be defined in section 4.4. Conceptually, they are the similar to the terms applied to data.)

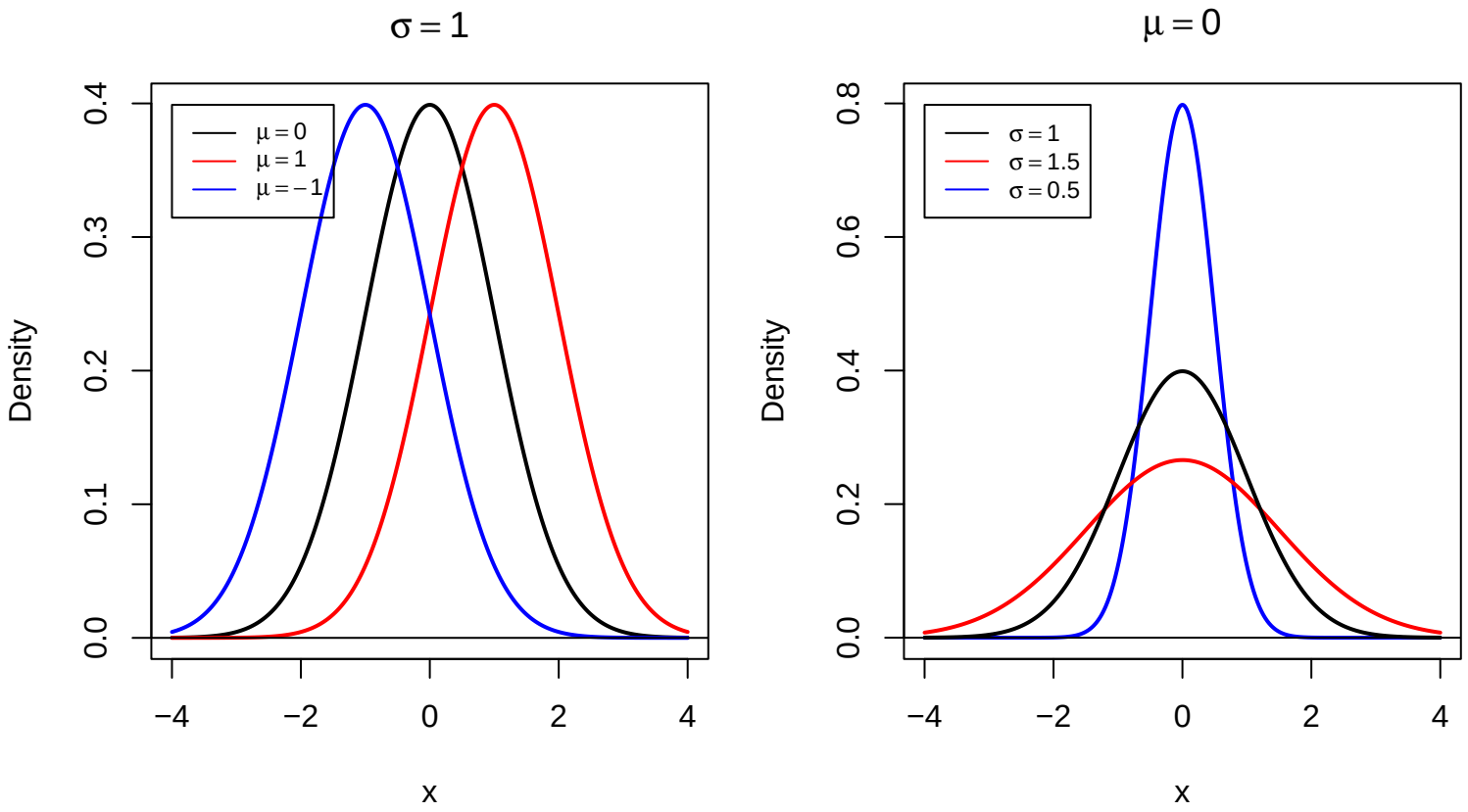
The basic form of the normal is:



The PDF for the for the normal distribution ($N(\mu, \sigma^2)$) is

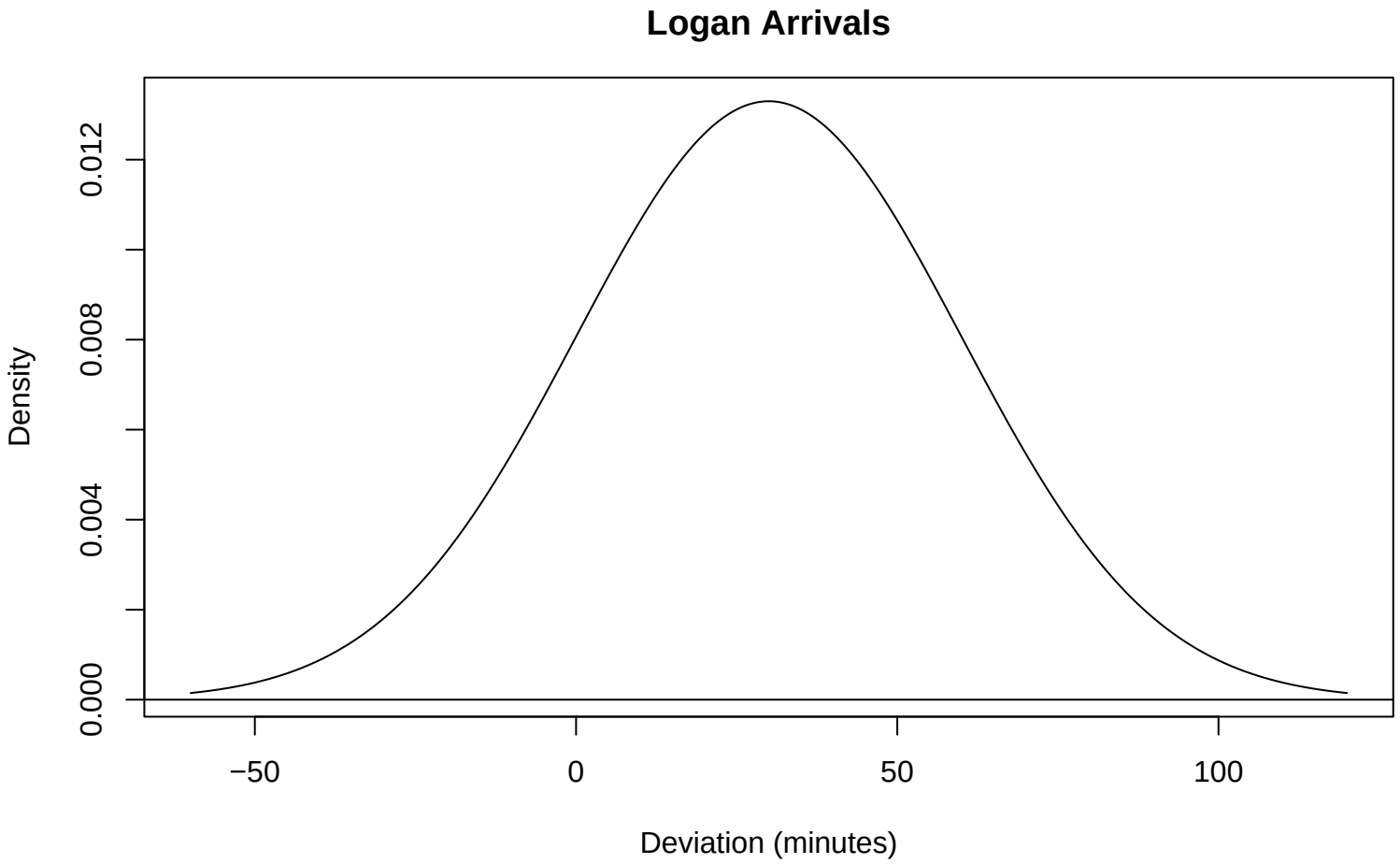
$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} e^{-(x-\mu)^2/2\sigma^2}$$

The mean μ describes where the density is centered and the standard deviation σ describes how spread out the density is.



The normal is symmetric where the median = mean. (For a density curve, the median is the value that has area exactly 0.5 to the left of it.)

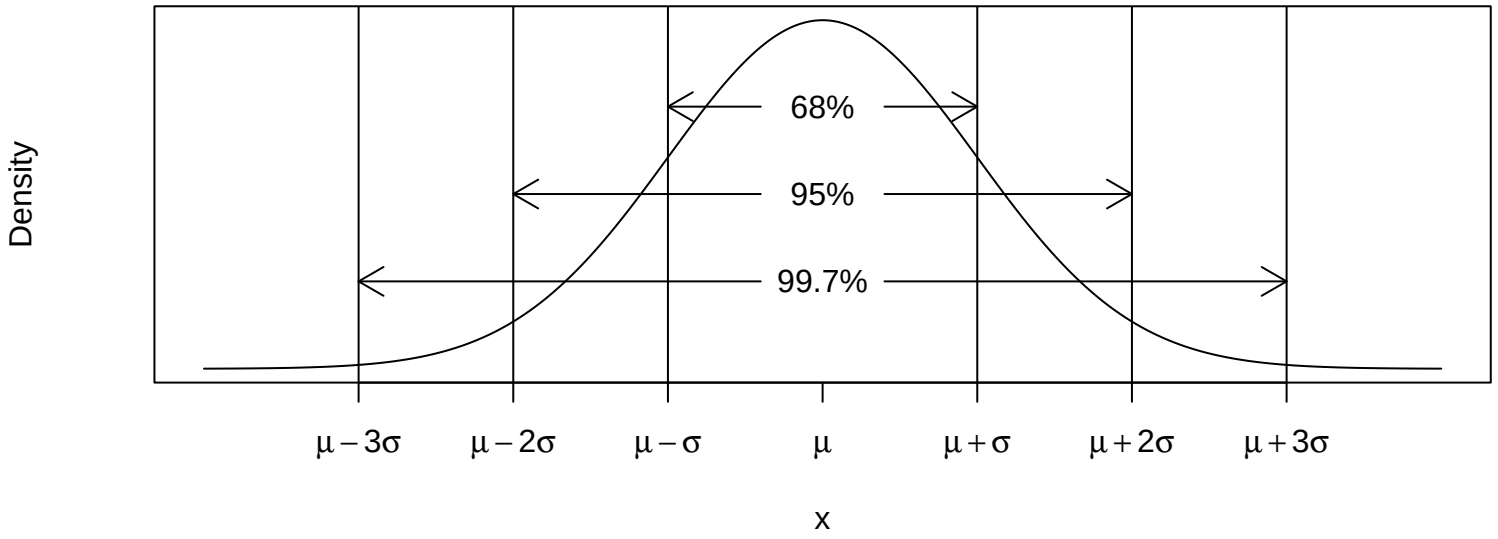
The Logan arrival times example from last class was a normal distribution with $\mu = 30$ and $\sigma = 30$.



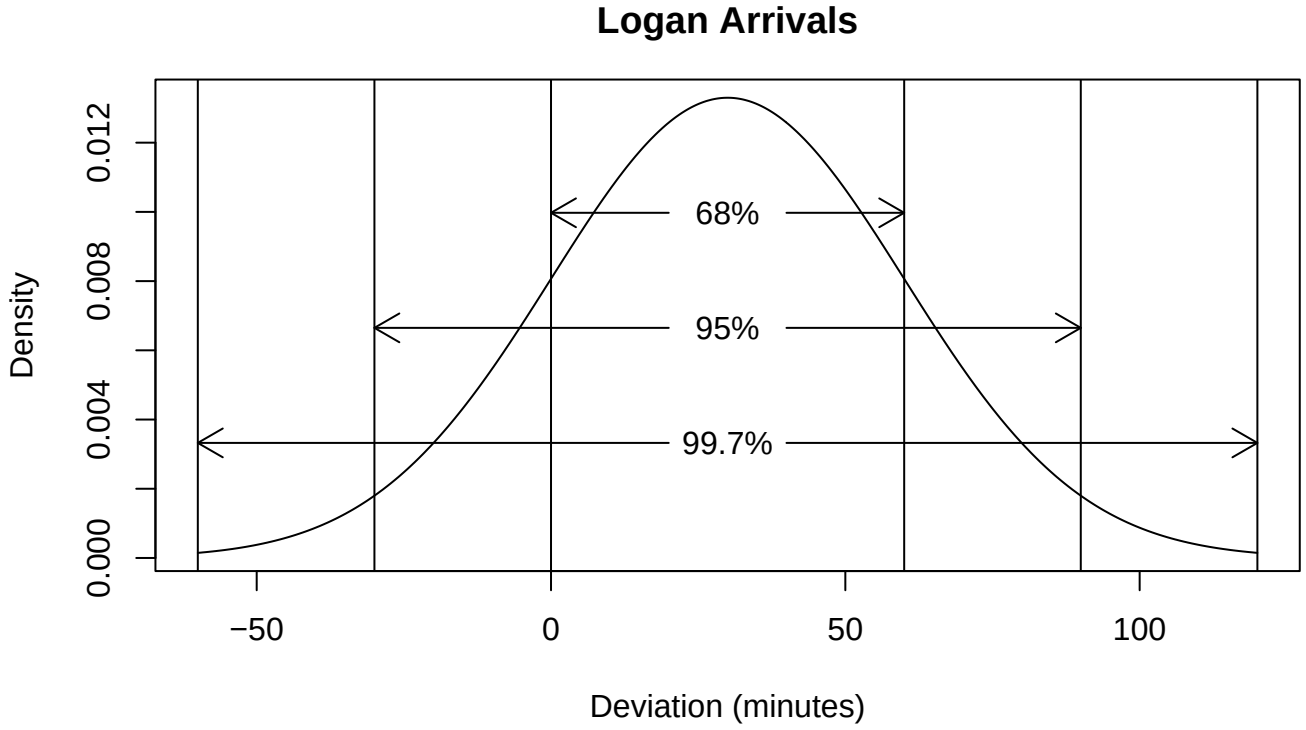
68-95-99.7 Rule

For a normal with distribution with mean μ and standard deviation σ,

- Approximately 68% of the area falls within 1σ of the mean μ
- Approximately 95% of the area falls within 2σ of the mean μ
- Approximately 99.7% of the area falls within 3σ of the mean μ



So assuming the model describing the arrival times at Logan is accurate (I'm sure its not), this rule suggests that about 68% of planes will arrive just on time to 1 hour late, 95% will arrive between 30 minutes early and 90 minutes late, and virtually all will be between 1 hour early and 2 hours late.



The 68-95-99.7 rule suggests that in some sense all normal distributions are the same. In fact, this can be shown if we measure in units of size σ about the mean μ.

Changing to these units is known as standardization

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

These standardized values are often called z-scores. A z-score tells us how many standard deviations the original value is from its mean.

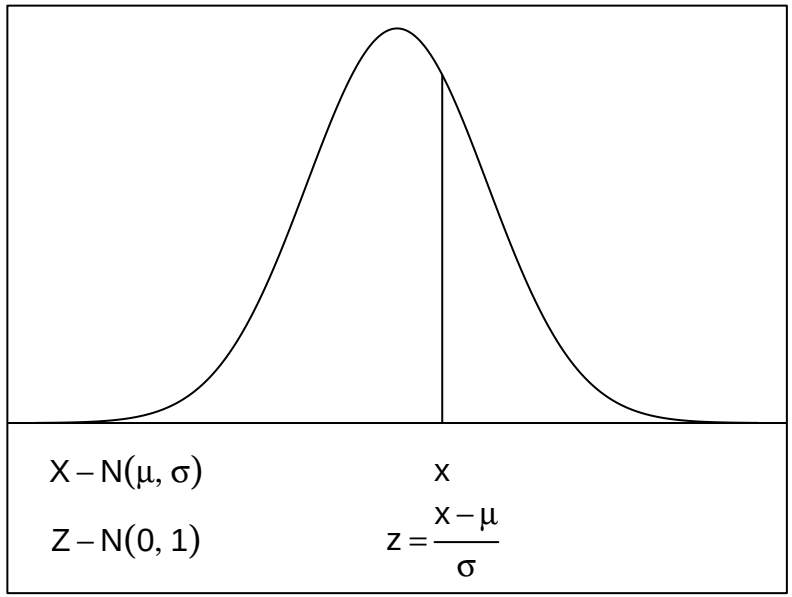
Standard Normal Distribution

The normal distribution with mean 0 and standard deviation 1 $(N(0,1))$

If a variable X has a normal distribution $N(\mu, \sigma)$, then the standardized variable

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

is standard normal.



Calculating Normal Probabilities

Since any normal distribution can be transformed to a standard normal, we only need to know how to get probabilities for the standard normal.

Standard Normal Table

Inside the front cover of IPS is a standard normal table. It gives probabilities of the form $P[Z \leq z]$.

Rows: first two digits of the z-score.

Column: the third digit of the z-score.

So for example, to get $P[Z \leq 1.28]$, go to row 1.2 and column 0.08, which gives

$$P[Z \leq 1.28] = 0.8997$$

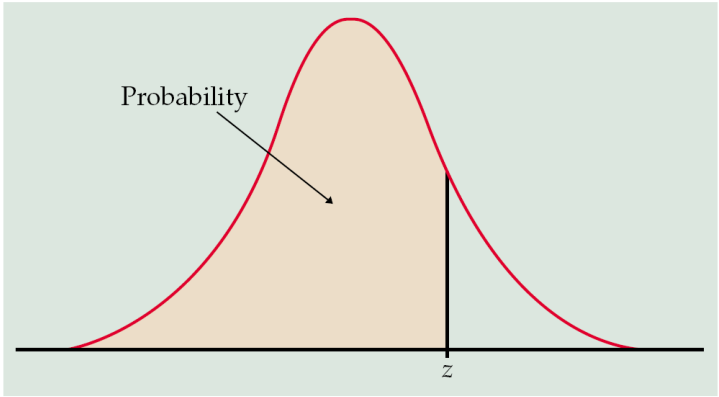


Table entry for z is the area under the standard normal curve to the left of z.

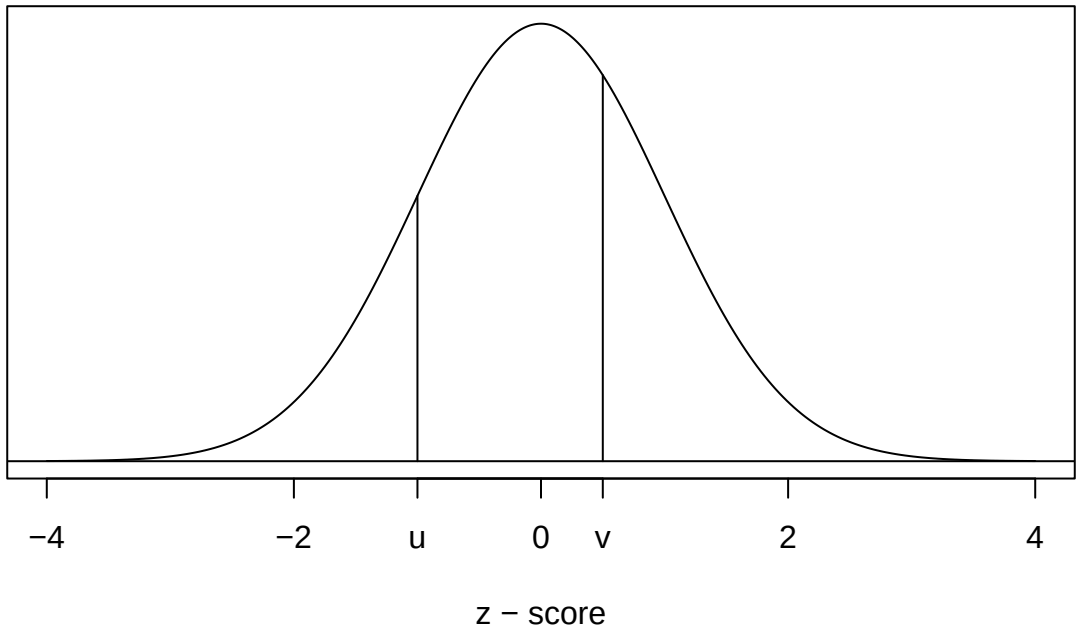
TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

Using this table, we can get any probability involving standard normals, which implies we can get any probability involving normals.

1. $P[Z > v] = 1 - P[Z \leq v]$

2. $P[u \leq Z \leq v] = P[Z \leq v] - P[Z \leq u]$



So for example

$$P[Z > 1.50]$$
$$= 1 - P[Z \leq 1.50]$$
$$= 1 - 0.9332 = 0.0668$$

$$P[1 \leq Z \leq 2]$$
$$= P[Z \leq 2] - P[Z \leq 1]$$
$$= 0.9972 - 0.8413$$
$$= 0.1359$$

Table entry for z is the area under the standard normal curve to the left of z.

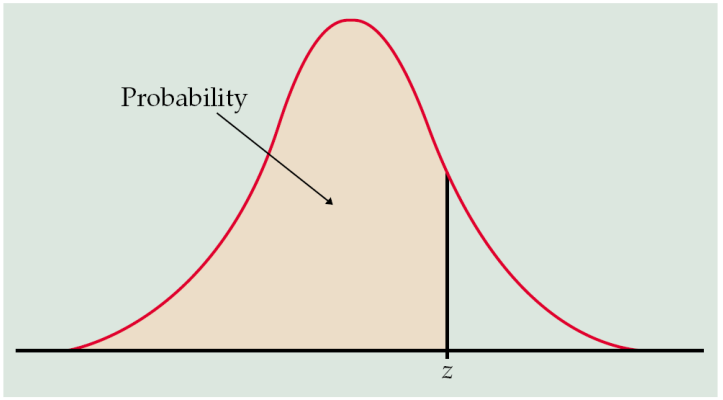


TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

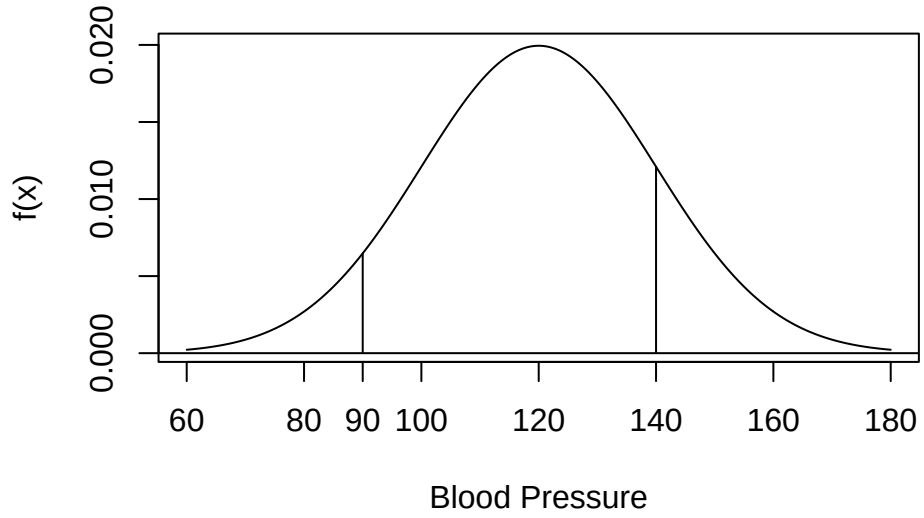
Suppose for example that blood pressure (X) can be modelled (approximately) by a normal distribution with $\mu = 120$ and $\sigma = 20$.

Suppose we are interested in the following probabilities

1) $P[X \leq 90]$

2) $P[X > 140]$

3) $P[90 \leq X \leq 140]$



To calculate these probabilities we need to preform the following steps

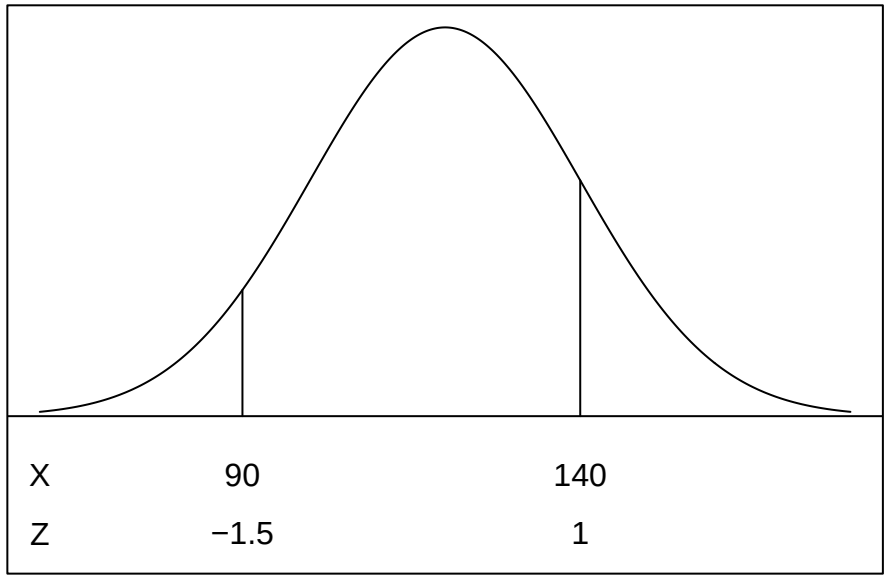
1. Standardize the necessary endpoints.

2. Get probabilities from the table

3. Do necessary arithmetic

1) $P[X \leq 90]$

$$z(90) = \frac{90 - 120}{20} = -1.5$$



$$P[X \leq 90] = P[Z \leq 1.5] = 0.0668$$

2) $P[X > 140]$

$$z(140) = \frac{140 - 120}{20} = 1$$

$$P[X > 140] = P[Z > 1] = 1 - P[Z < 1] = 1 - 0.8413 = 0.1587$$

3) $P[90 \leq X \leq 140]$

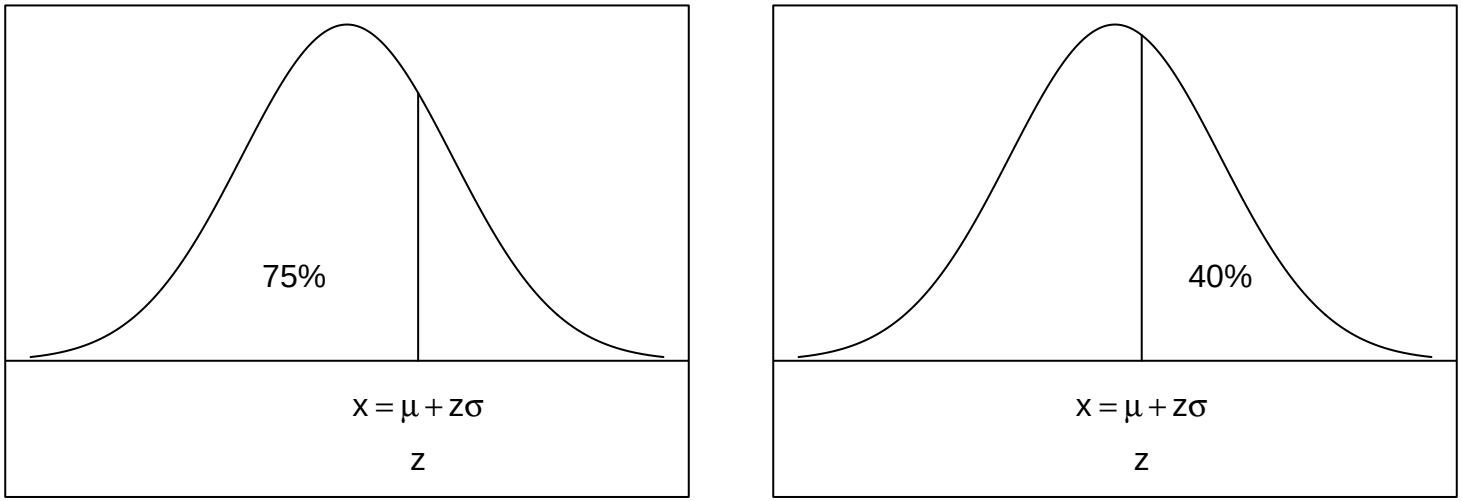
$$P[90 \leq X \leq 140] = P[-1.5 \leq Z \leq 1] = 0.8413 - 0.0668 = 0.7745$$

Inverse Problem

Find values of the random variable that correspond to probabilities of interest.

1. 75% of people have blood pressures less than what?

2. What blood pressure is exceeded by 40% of people?



If $Z \sim N(0,1)$ then $X = \sigma Z + \mu \sim N(\mu, \sigma)$

Steps:

1. Find percentile z^* for the standard normal

2. Transform to the desired units

$$x^* = \sigma z^* + \mu$$

To do step 1), find the probability in the table closest to the desired probability. Then the row and column give z^*.

1) 75% of people have a blood pressure less than what?

What z^* gives
$P[Z \leq z^*] \approx 0.75$

$$z^* = 0.67$$

Now transform back to original units

$$\begin{aligned} x^* &= 20 \times z^* + 120 \\ &= 20 \times 0.67 + 120 \\ &= 133.4 \end{aligned}$$

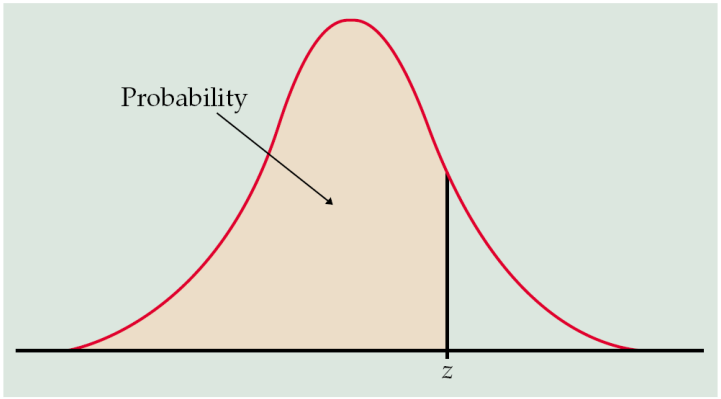


Table entry for z is the area under the standard normal curve to the left of z.

TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

2) What blood pressure is exceeded by 40% of people?

What z^* gives
$P[Z \geq z^*] \approx 0.40$

This is equivalent to
$P[Z \leq z^*] \approx 0.60$

$$z^* = 0.25$$

$$\begin{aligned} x^* &= 20 \times z^* + 120 \\ &= 20 \times 0.25 + 120 \\ &= 125 \end{aligned}$$

Table entry for z is the area under the standard normal curve to the left of z.

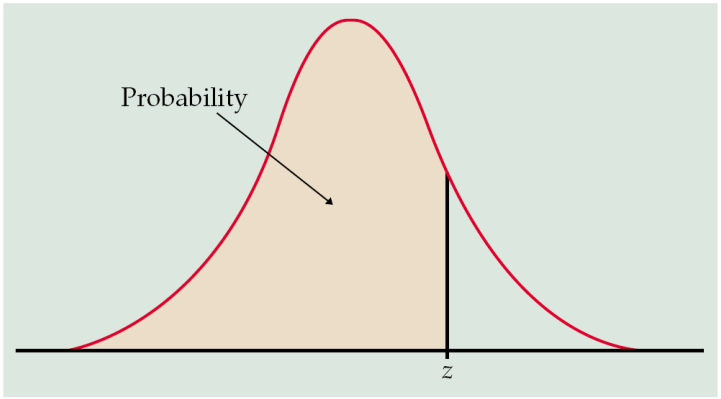


TABLE A Standard normal probabilities (*continued*)

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	.5000	.5040	.5080	.5120	.5160	.5199	.5239	.5279	.5319	.5359
0.1	.5398	.5438	.5478	.5517	.5557	.5596	.5636	.5675	.5714	.5753
0.2	.5793	.5832	.5871	.5910	.5948	.5987	.6026	.6064	.6103	.6141
0.3	.6179	.6217	.6255	.6293	.6331	.6368	.6406	.6443	.6480	.6517
0.4	.6554	.6591	.6628	.6664	.6700	.6736	.6772	.6808	.6844	.6879
0.5	.6915	.6950	.6985	.7019	.7054	.7088	.7123	.7157	.7190	.7224
0.6	.7257	.7291	.7324	.7357	.7389	.7422	.7454	.7486	.7517	.7549
0.7	.7580	.7611	.7642	.7673	.7704	.7734	.7764	.7794	.7823	.7852
0.8	.7881	.7910	.7939	.7967	.7995	.8023	.8051	.8078	.8106	.8133
0.9	.8159	.8186	.8212	.8238	.8264	.8289	.8315	.8340	.8365	.8389
1.0	.8413	.8438	.8461	.8485	.8508	.8531	.8554	.8577	.8599	.8621
1.1	.8643	.8665	.8686	.8708	.8729	.8749	.8770	.8790	.8810	.8830
1.2	.8849	.8869	.8888	.8907	.8925	.8944	.8962	.8980	.8997	.9015
1.3	.9032	.9049	.9066	.9082	.9099	.9115	.9131	.9147	.9162	.9177
1.4	.9192	.9207	.9222	.9236	.9251	.9265	.9279	.9292	.9306	.9319
1.5	.9332	.9345	.9357	.9370	.9382	.9394	.9406	.9418	.9429	.9441
1.6	.9452	.9463	.9474	.9484	.9495	.9505	.9515	.9525	.9535	.9545
1.7	.9554	.9564	.9573	.9582	.9591	.9599	.9608	.9616	.9625	.9633
1.8	.9641	.9649	.9656	.9664	.9671	.9678	.9686	.9693	.9699	.9706
1.9	.9713	.9719	.9726	.9732	.9738	.9744	.9750	.9756	.9761	.9767
2.0	.9772	.9778	.9783	.9788	.9793	.9798	.9803	.9808	.9812	.9817
2.1	.9821	.9826	.9830	.9834	.9838	.9842	.9846	.9850	.9854	.9857
2.2	.9861	.9864	.9868	.9871	.9875	.9878	.9881	.9884	.9887	.9890
2.3	.9893	.9896	.9898	.9901	.9904	.9906	.9909	.9911	.9913	.9916
2.4	.9918	.9920	.9922	.9925	.9927	.9929	.9931	.9932	.9934	.9936

Assessing Normality

One of the reasons mentioned earlier for examining normal curves is that they are often good descriptions for real data.

May want to check that assumption when dealing with data.

1. Does the histogram look normal

2. Does data match the 68-95-99.7 rule

3. Normal Quantile plots (Also called Normal Probability plots or Normal Scores plots

The idea behind Normal Quantile plots: Are the ordered values spread out properly?

Compares the ordered data with what would be expected if the data were really normal.

If the data is approximately normal, the points on a Normal Quantile plot should lie close to a straight line. Systematic deviations from a straight line indicate non-normal data. Outliers appear as points that are far away from the overall pattern.

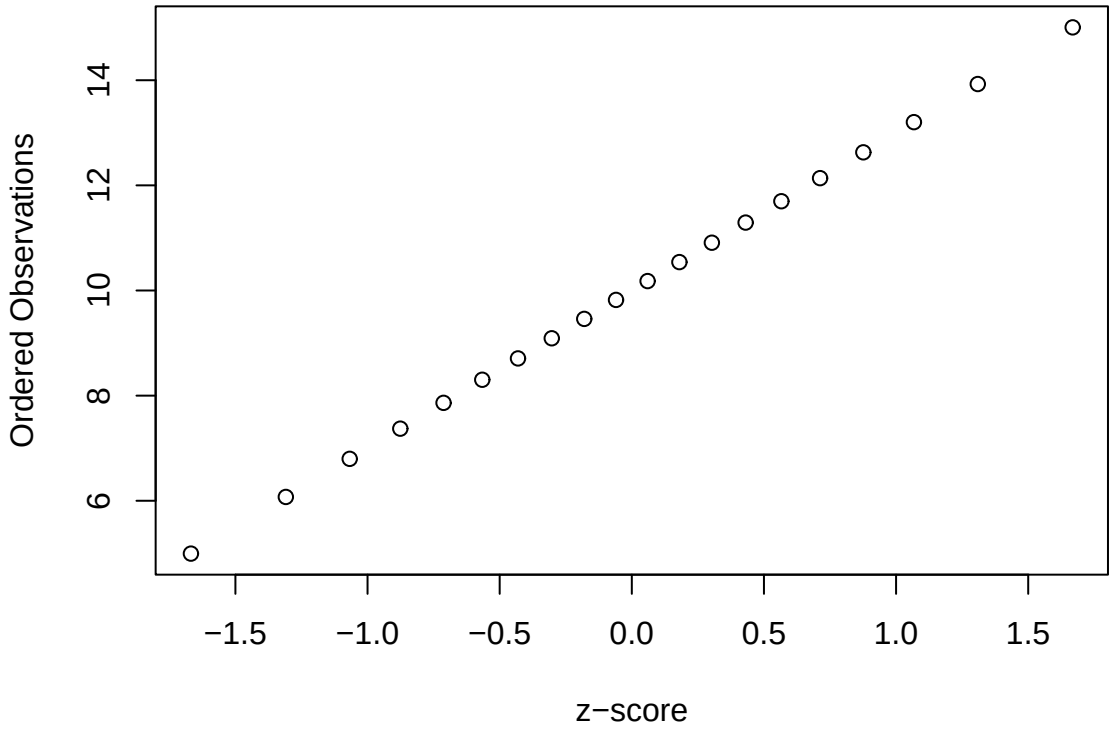
Assume $y_{(1)} \leq y_{(2)} \leq \ldots \leq y_{(n)}$ are the ordered data values and let $z_1 \leq z_2 \leq \ldots z_n$ be the expected order statistics from a standard normal. z_k approximately satisfies

$$P[Z \leq z_k] = \frac{k + 0.5}{n + 1}; \qquad k = 1, 2, \ldots, n$$

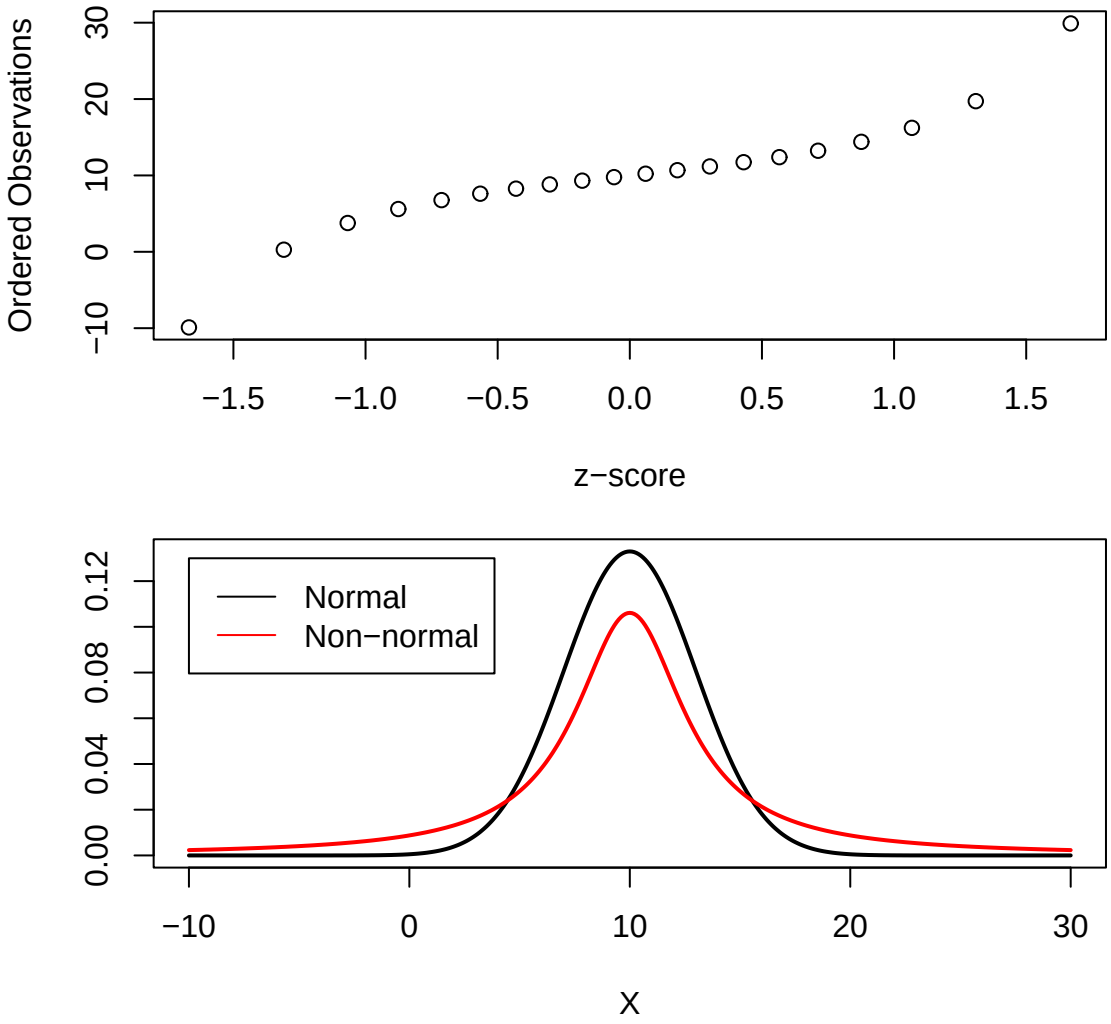
The Normal Quantile plots z_k on the x-axis vs $y_{(k)}$ on the y-axis.

Idealized patterns

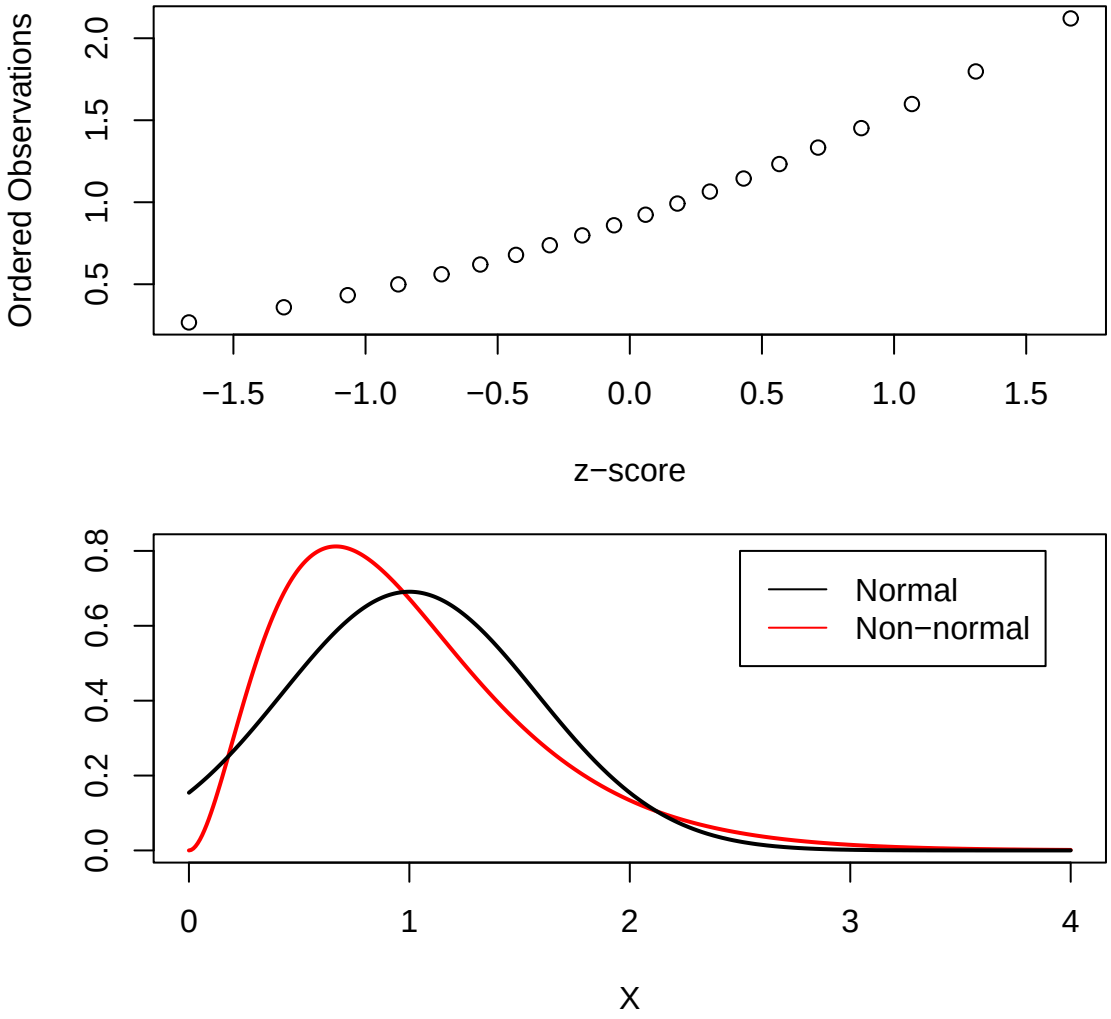
- Desirable



- Overdispersed (long tailed)

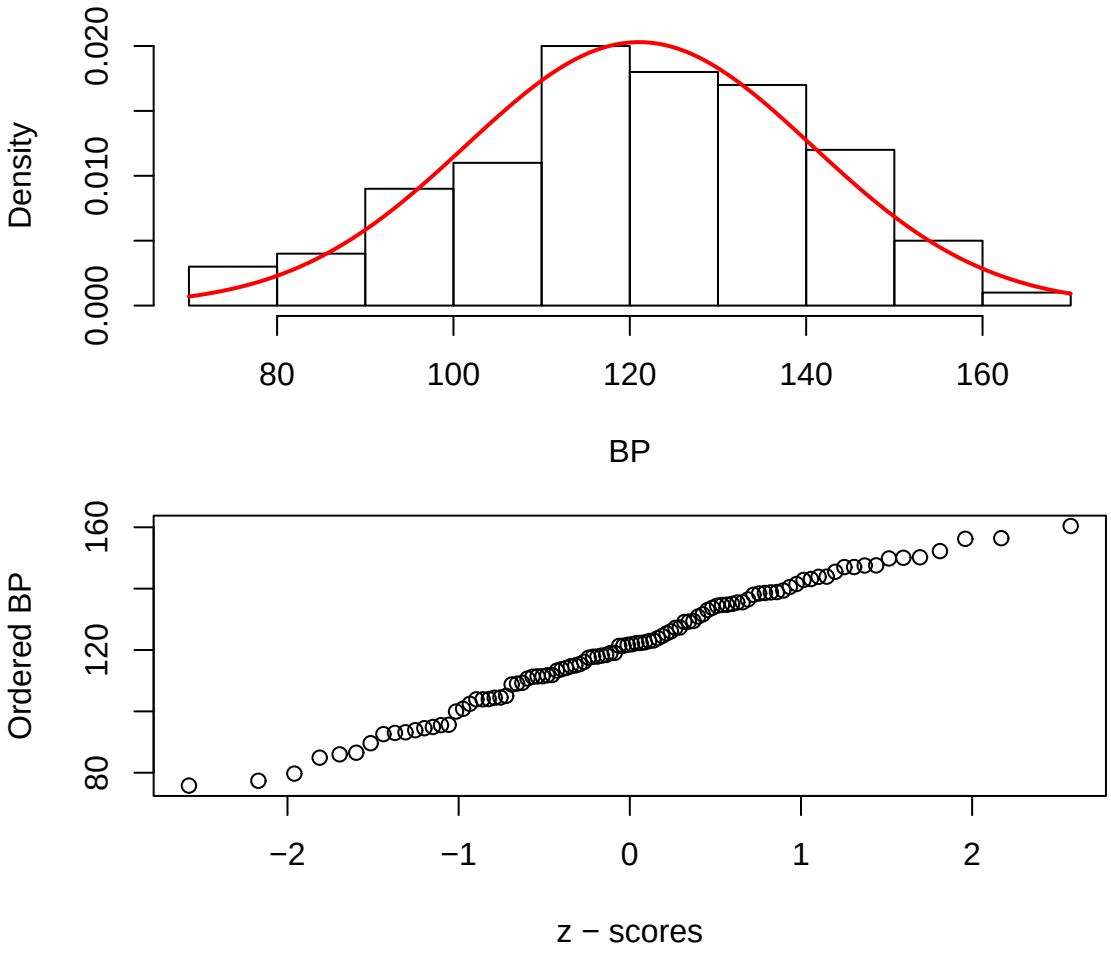


- Skewed right

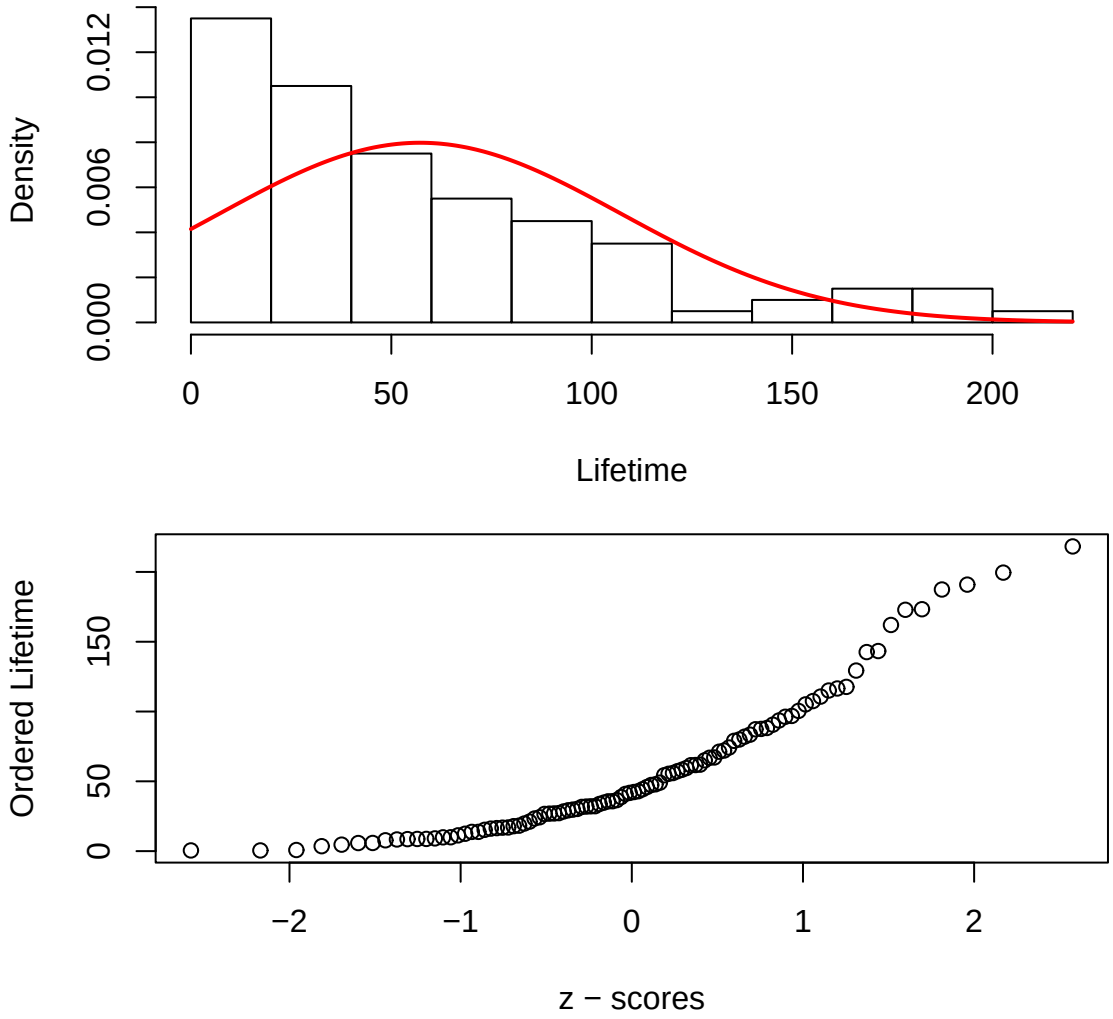


Simulated Examples

- Simulated Normal Data

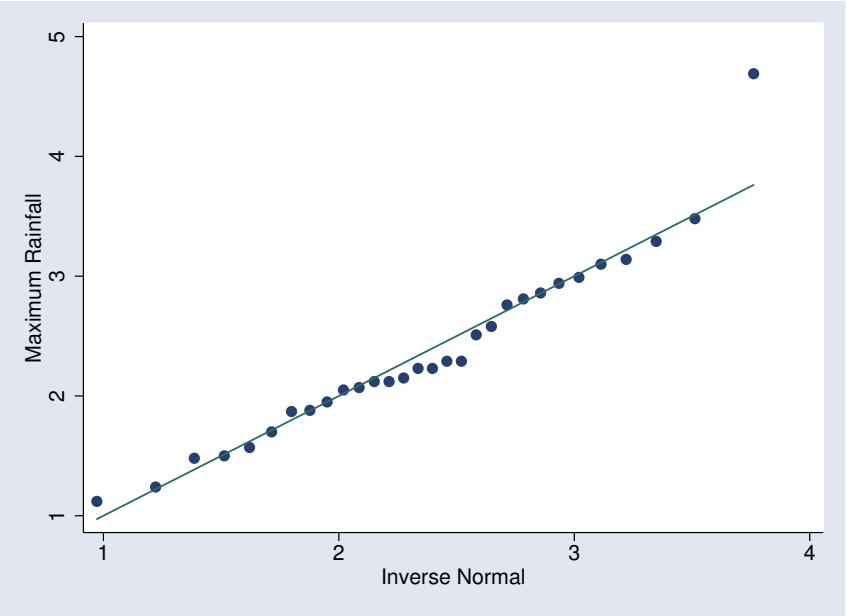


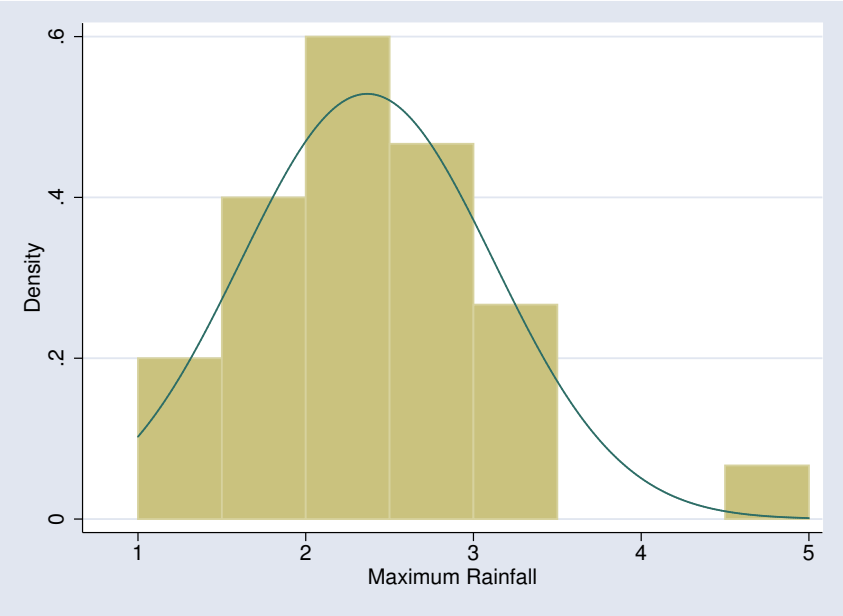
- Simulated Skewed Data



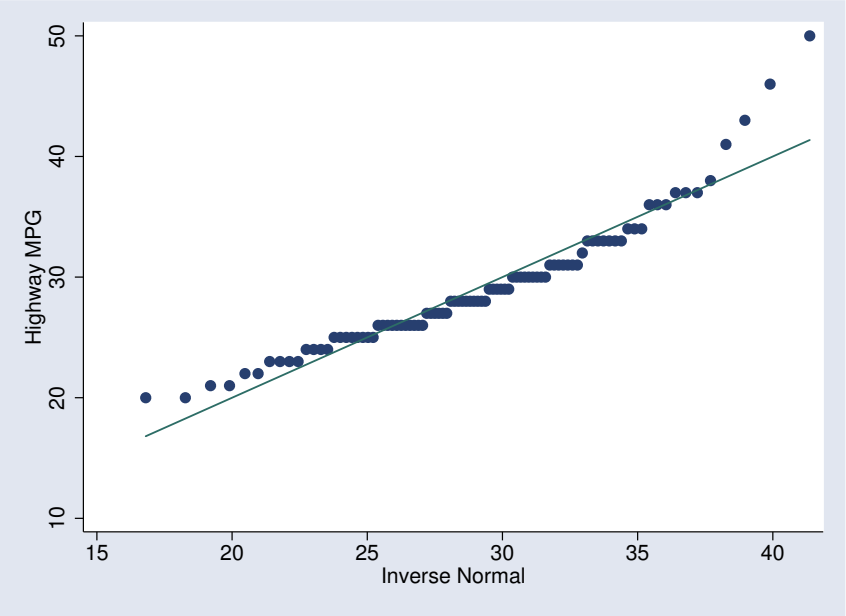
Real Examples

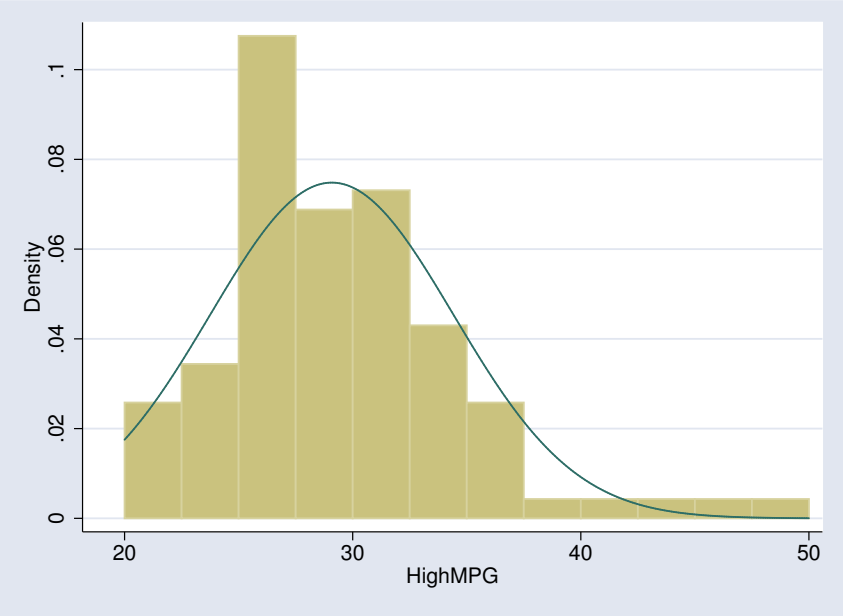
- South Bend Rainfall





- Highway MPG





Note: In Stata, the Normal Quantile (which is available under `Graphics > Distributional Plots` menu or with the `qnorm` command) plot the ordered data against the expected order statistics for a normal with the same mean and standard deviation as the data.

It plots $sz_k + \bar{y}$ vs $y_{(k)}$. This doesn't change the use of the graph as all it is doing is relabelling the x-axis in the plot