W1211 Introduction to Statistics
Lecture 11

Wei Wang

Oct 10th, 2012

Uniform RV

- We call a uniform rv U a standard uniform, if and only if U ~ uniform on [0,1]

- For a standard uniform rv U, we can easily calculate,

$$\mathrm{E}(\mathrm{U}) = \int_0^1 x \cdot 1 dx = \frac{1}{2}$$

$$\mathrm{E}(\mathrm{U}^2) = \int_0^1 x^2 \cdot 1 dx = \frac{1}{3}$$

$$\mathrm{Var}(\mathrm{U}) = \mathrm{E}(\mathrm{U}^2) - [\mathrm{E}(\mathrm{U})]^2 = \frac{1}{12}$$

General Uniform

- Note that a general case of uniform distribution X on [*A, B*] can be treated as a linear transform of a standard uniform, i.e., $X = (B - A)U + A$.

- Proposition:

> If X is a continuous uniform rv on [*A, B*], then
> $E(X) = (B + A)/2$, $Var(X) = (B - A)^2/12$

- R command:

```
dunif(x, min=0, max=1),
punif(q, min=0, max=1),
qunif(p, min=0, max=1).
```

The Normal Distribution

- It's probably the most important distribution in the world!

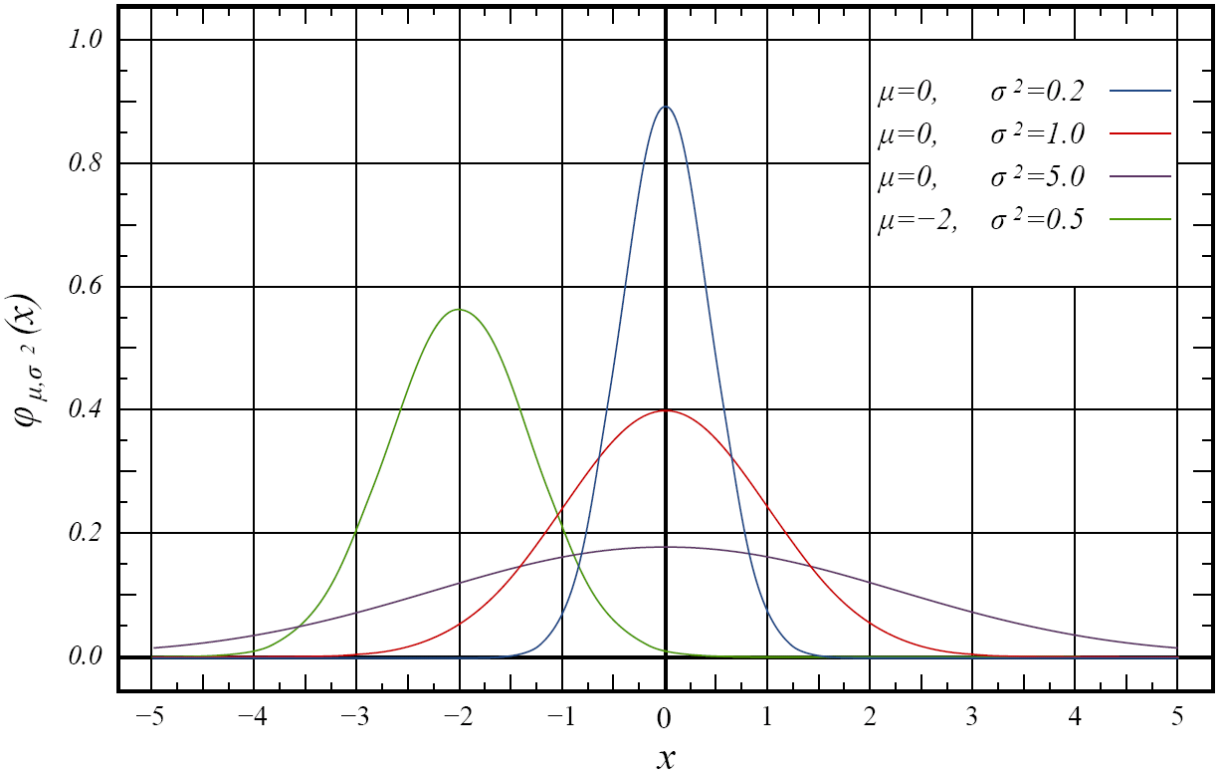
- Many numerical populations have distributions that can be fit very closely by an appropriate normal curve. (people's height/weight; testing scores; etc.) Even when the underlying distribution is discrete, (yearly number of customers to Wal-Mart; etc.) the normal curve often gives an excellent approximation.

- A continuous rv is said to have a normal (Gaussian) distribution with parameters μ and σ, where $-\infty<\mu<\infty$, and $0<\sigma$, if the pdf of X is

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

The Normal pdf

- Normal distribution is a bell-shaped, single peaked and symmetric distribution.



Parameters

- Clearly $f(x; \mu, \sigma) \geq 0$, but a somewhat complicated calculus argument must be used to verify that

$$\int_{-\infty}^{\infty} f(x; \mu, \sigma) dx = 1.$$

- Parameter μ, stands for the expected value of the normal distribution.

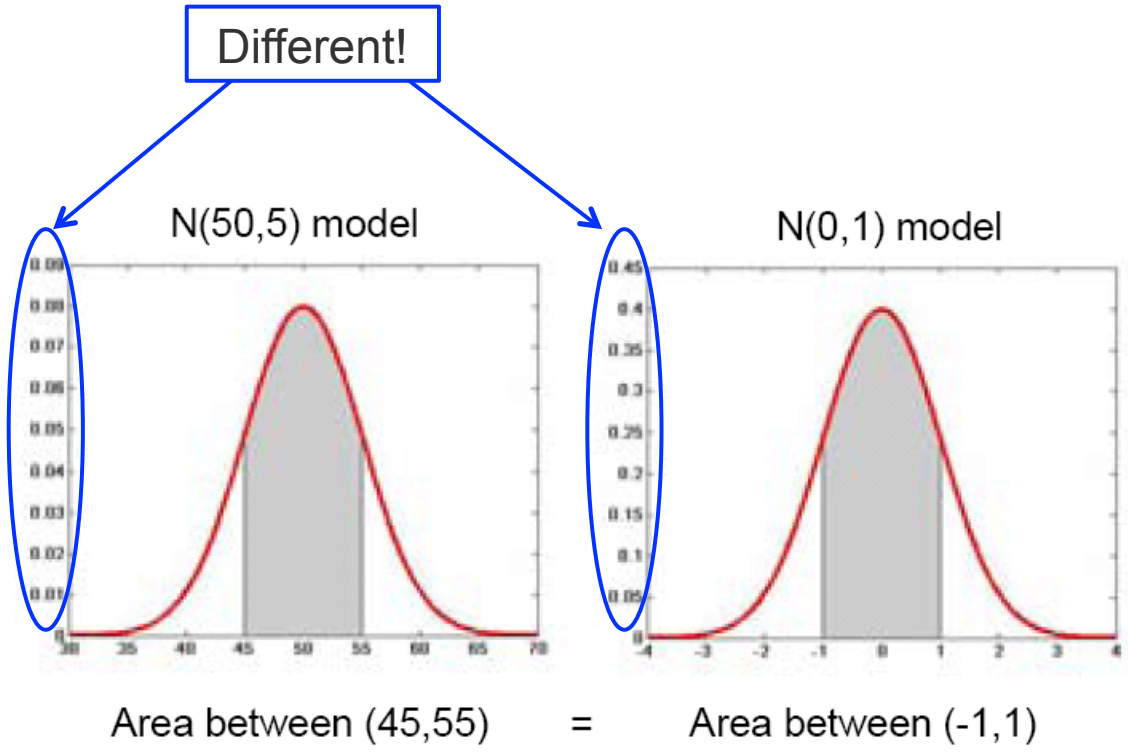
  Exercise: show that if $X \sim N(\mu, \sigma^2)$, then $E(X) = \mu$.

- Parameter σ, stands for the standard deviation of the normal distribution.

  Exercise: show that if $X \sim N(\mu, \sigma^2)$, then $Var(X) = \sigma^2$.

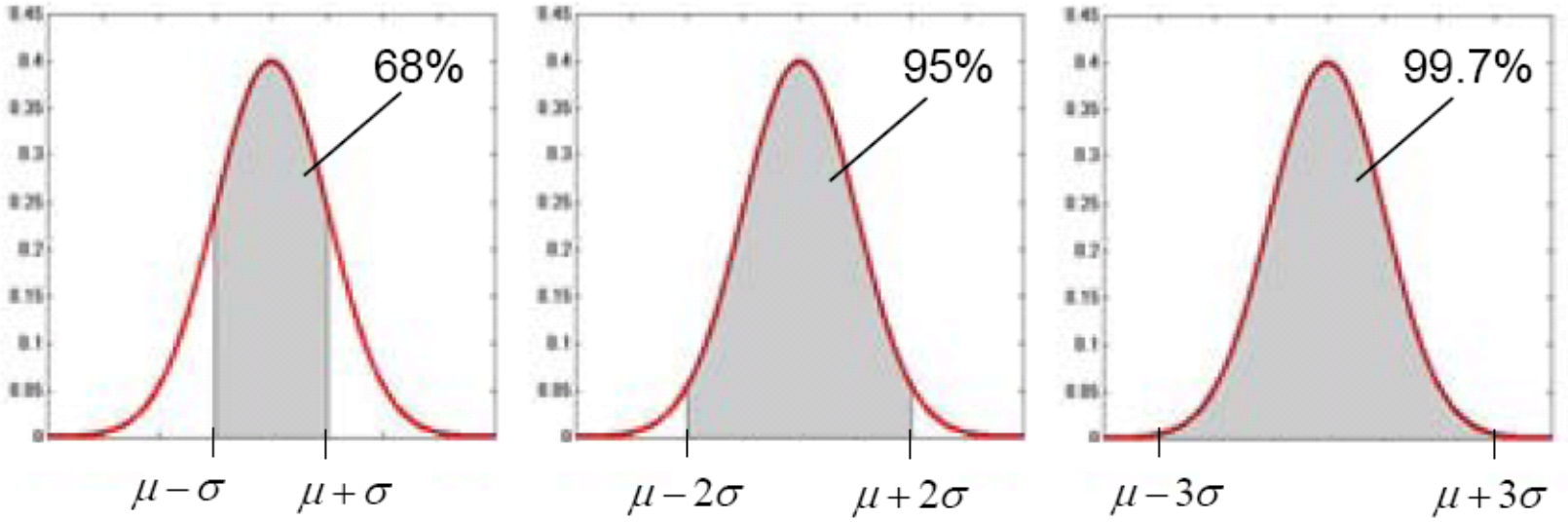
Basic Properties

- All normal models have the same shape and the same area within *x* standard deviations of its mean.



The 68-95-99.7 Rule

- For any normal distribution, we have the following result:



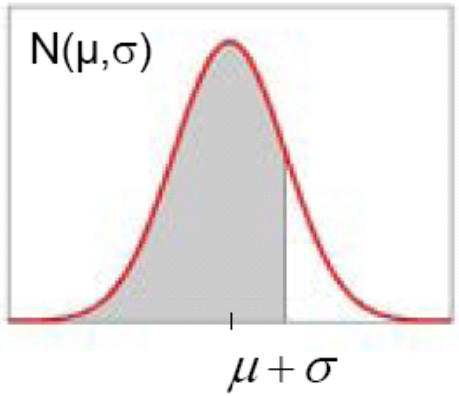
Example

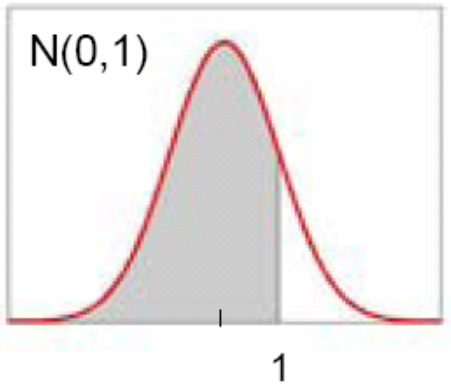
Ex. On an exam the scores followed an approximate normal model with $\mu = 72$ and $\sigma = 8$.

- 68% of the students scored between 72±8 or (64, 80).
- 95% of the scores were between 72±2*8 or (56, 88).
- 99.7% of the scores were between 72±3*8 or (48, 96).

- What proportion scored below 84?

Key Result





$$area\{y < \mu+\sigma\} \quad = \quad area\{z < 1\}$$

Standard Normal

- If $Z \sim N(0, 1)$, i.e., if Z is a normal random variable with $\mu=0$, $\sigma=1$. Then Z is said to have a standard normal distribution.

- Any normally distributed rv's could be obtained by using standard normal rv's. To put it more mathematically, if $X \sim N(\mu, \sigma^2)$, then X could be written as

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

  where Z is a standard normal rv.

- Conversely, if $X \sim N(\mu, \sigma^2)$, then

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

  has a standard normal distribution. And Z is often called the "*z-score*" of X.

Example cont.

Ex. The exam scores followed a N(72,8) model.

What proportion of the students scored below 84?

$$z = \frac{y - \mu}{\sigma} = \frac{84 - 72}{8} = 1.5$$

Answer: 93.32%

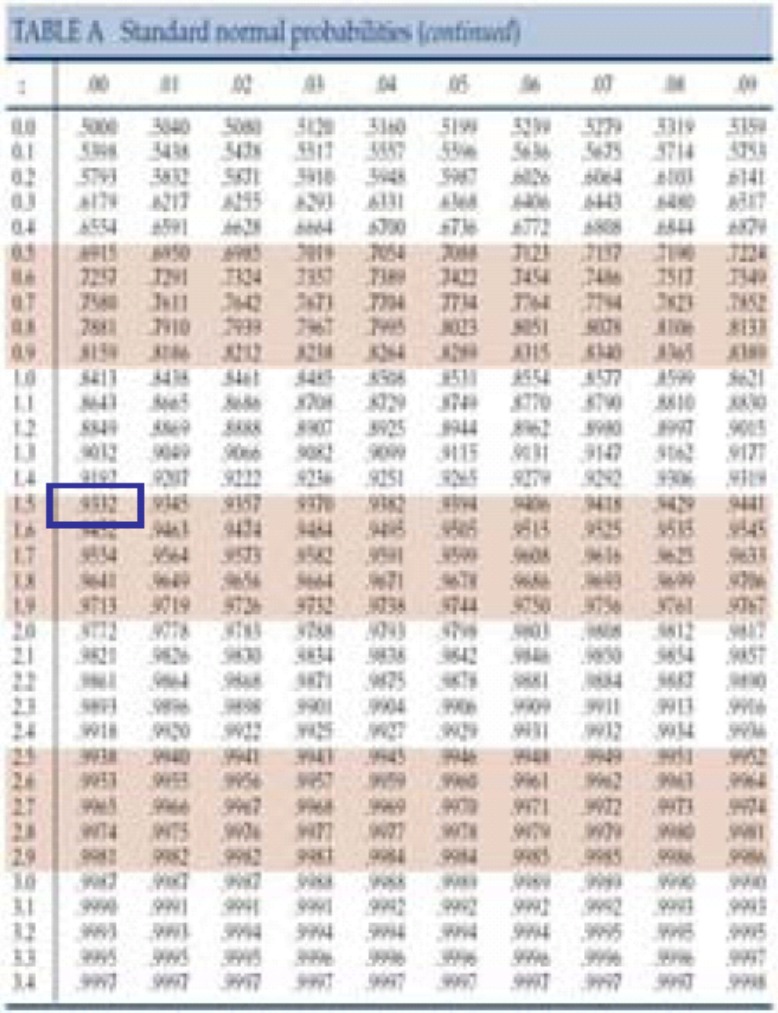


TABLE A Standard normal probabilities (continued)

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
2.5	.9938	.9940	.9941	.9943	.9945	.9946	.9948	.9949	.9951	.9952
2.6	.9953	.9955	.9956	.9957	.9959	.9960	.9961	.9962	.9963	.9964
2.7	.9965	.9966	.9967	.9968	.9969	.9970	.9971	.9972	.9973	.9974
2.8	.9974	.9975	.9976	.9977	.9977	.9978	.9979	.9979	.9980	.9981
2.9	.9981	.9982	.9982	.9983	.9984	.9984	.9985	.9985	.9986	.9986
3.0	.9987	.9987	.9987	.9988	.9988	.9989	.9989	.9989	.9990	.9990
3.1	.9990	.9991	.9991	.9991	.9992	.9992	.9992	.9992	.9993	.9993
3.2	.9993	.9993	.9994	.9994	.9994	.9994	.9994	.9995	.9995	.9995
3.3	.9995	.9995	.9995	.9996	.9996	.9996	.9996	.9996	.9996	.9997
3.4	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9997	.9998

Simplification

- Thus, any problem about any normal rv $X \sim N(\mu, \sigma^2)$, can be translated to a problem about a standard normal rv Z.

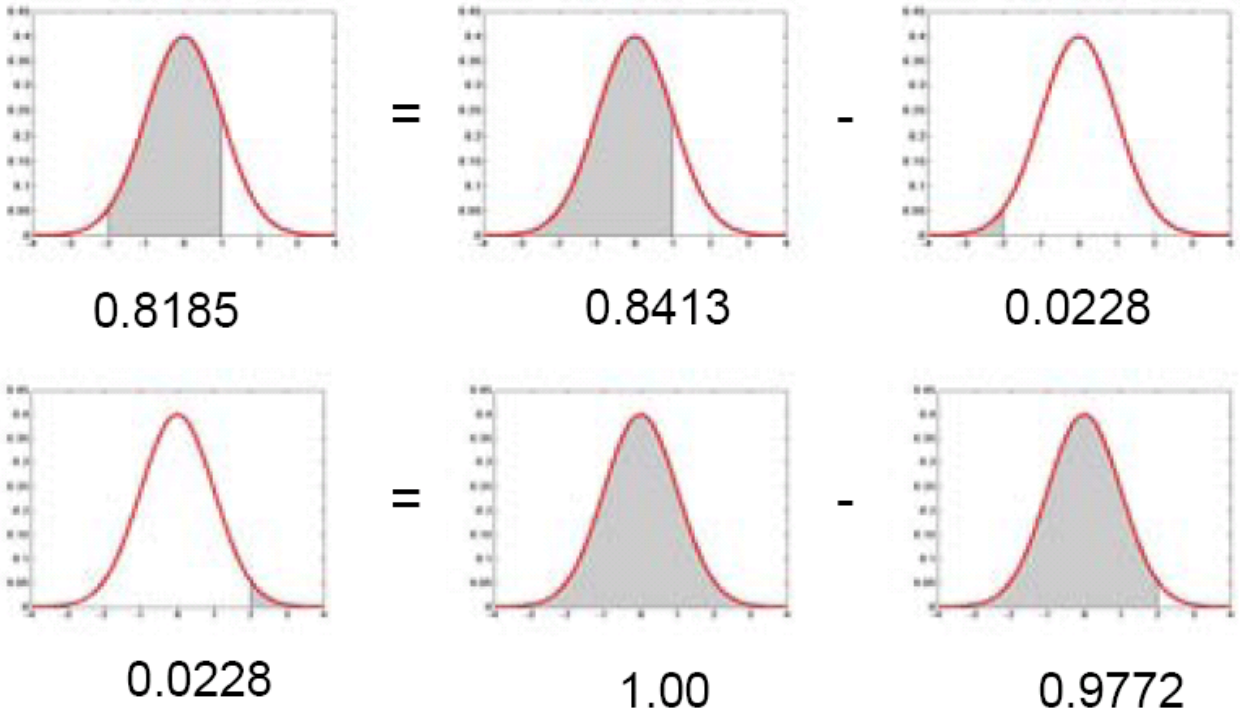
Ex. $P(a \le X \le b) = P[(a-\mu)/\sigma \le (X-\mu)/\sigma \le (b-\mu)/\sigma] = P[(a-\mu)/\sigma \le Z \le (b-\mu)/\sigma]$.

- The cumulative distribution function of standard normal distribution, that is $\Phi(z) = P(Z \le z)$, is already known! (Appendix Table.)

- Check Table A.3 to determine $P(Z \le 0.76)$; $P(Z > 0.76)$; $P(-1.32 \le Z \le 0.76)$.

- Question: How to get the p-th percentile of the standard normal from A.3?

Using the Normal Table



R instead of tables

- R command:
```
dnorm(x, mean = 0, sd = 1),
pnorm(q, mean = 0, sd = 1),
qnorm(p, mean = 0, sd = 1).
```

Example

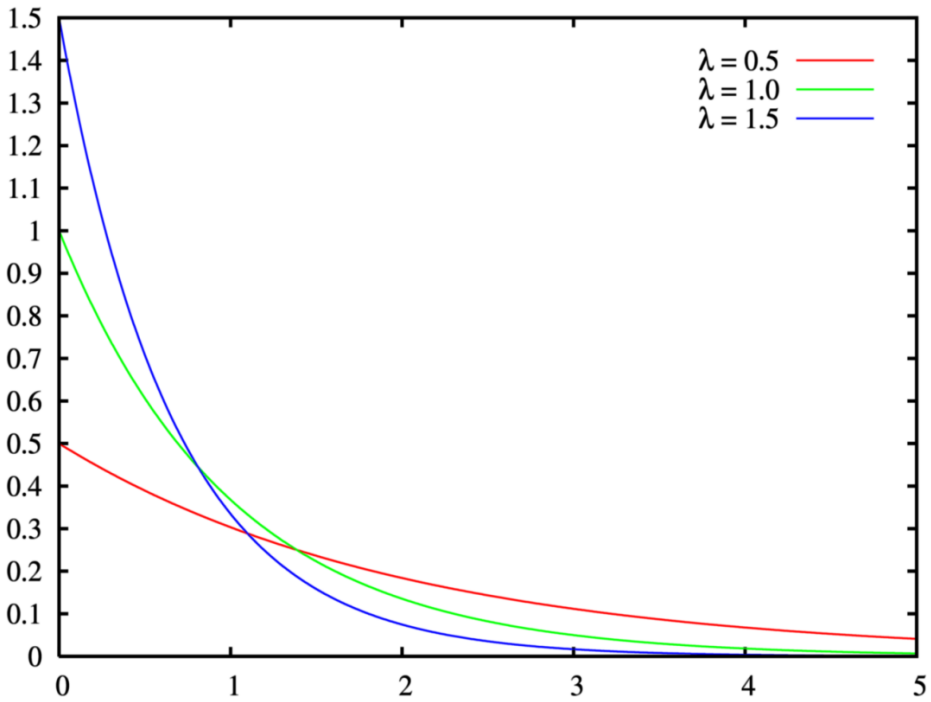
Ex. Suppose the height of all Columbia students can be described by a N(68, 4) model.

1. What proportion of students is shorter than 74 inches?
2. What proportion of students is taller than 74 inches?
3. How tall does a student have to be to be among the 10% tallest students?

The Exponential Distribution

- X is said to have an exponential distribution with parameter λ (λ>0) if the pdf of X is

$$f(x;\lambda) = \begin{cases} \lambda e^{-\lambda x} & x \geq 0 \\ 0 & \text{otherwise} \end{cases}$$



More on Exponential

- Note that an exponential rv X can only take positive values. And the cdf of X is

$$F(x;\lambda) = \begin{cases} \int_0^x \lambda e^{-\lambda y} dy = 1 - e^{-\lambda x} & x \geq 0 \\ 0 & x < 0 \end{cases}$$

- Thus $P(X>x) = 1 - F(x; \lambda) = e^{-\lambda x}$

- Proposition: (proof?)

> If X is an exponential rv with parameter λ, then
> $E(X) = 1/\lambda$, $Var(X) = 1/\lambda^2$

- R command:

```
dexp(x, lamda=1),
pexp(q, lamda=1),
qexp(p, lamda=1).
```

Exponential Distribution and Poisson Distribution

- Suppose the number of events occurring in a time interval of length t has Poisson Distribution with parameter αt, and the numbers of occurrences in non-overlapping intervals are independent of one another. Then the distribution of elapsed time between the occurrence of two successive events is exponential with parameter $\lambda = \alpha$.

Example

Ex. Suppose you are waiting for a bus at a bus station. And the distribution of the length of the time you have to wait to get on the bus after you arrive at the bus station is exponentially distributed with parameter λ. Assume you have already waited for s minutes, how much longer do you expect to wait?

First, we have to figure out the conditional probability distribution of the additional waiting time given we have waited for s minutes. For any $t > 0$

$$P(X \geq s + t \mid X \geq s) = P[(X \geq s + t) \cap (X \geq s)] / P(X \geq s)$$
$$= P(X \geq s + t)/P(X \geq s)$$
$$= e^{-\lambda t}$$

which is again an exponential distribution! Thus the expected additional waiting time is $1/\lambda$.

Memoryless Property

- From the previous example, we know that if a waiting time (or lifetime of something) follows an exponential distribution, *the distribution of additional waiting time* (*lifetime*) *is exactly the same as the original distribution of waiting time* (*lifetime*). In other words, the exponential distributed waiting time does NOT remember how much time you have waited, it starts afresh at any time!

- It is popular to model the distribution of component lifetime using the exponential distribution. However, the memoryless property may not be realistic in may applied problems. More general lifetime models can be furnished by the gamma, Weibull, and lognormal distributions. (Book: p159 – p168).

Normal Probability Plot

- Because of the important role that Normal Distribution plays in statistical inference, we often want to assess whether a given sample is roughly normal distributed. Normal Probability Plot is used for this purpose.

Normal Probability Plot

- Because of the important role that Normal Distribution plays in statistical inference, we often want to assess whether a given sample is roughly normal distributed. Normal Probability Plot is used for this purpose.
- The basic strategy is to compare sample features with population features. In probability plot, we are using sample percentile(quantile) and population percentile(quantile), so it is also known as Q-Q plot.

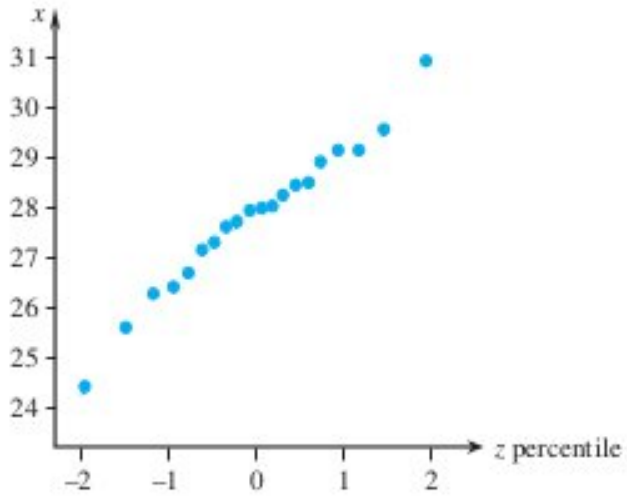
Normal Probability Plot

- Because of the important role that Normal Distribution plays in statistical inference, we often want to assess whether a given sample is roughly normal distributed. Normal Probability Plot is used for this purpose.
- The basic strategy is to compare sample features with population features. In probability plot, we are using sample percentile(quantile) and population percentile(quantile), so it is also known as Q-Q plot.
- The definition of a normal probability plot

> A plot of the n pairs
>
> ([100(i − .5)/n]th z percentile, ith smallest observation)
>
> on a two-dimensional coordinate system is called a **normal probability plot.** If the sample observations are in fact drawn from a normal distribution with mean value μ and standard deviation σ, the points should fall close to a straight line with slope σ and intercept μ. Thus a plot for which the points fall close to some straight line suggests that the assumption of a normal population distribution is plausible.

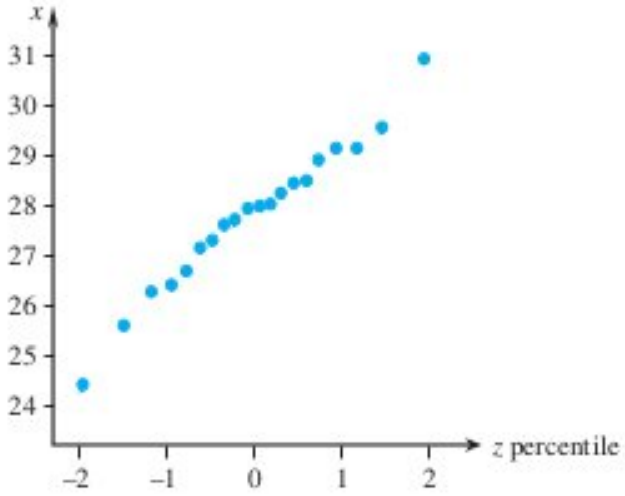
Examples of Normal Probability Plot

- A Normal Sample



Examples of Normal Probability Plot

- A Normal Sample



- Two Non-normal Samples

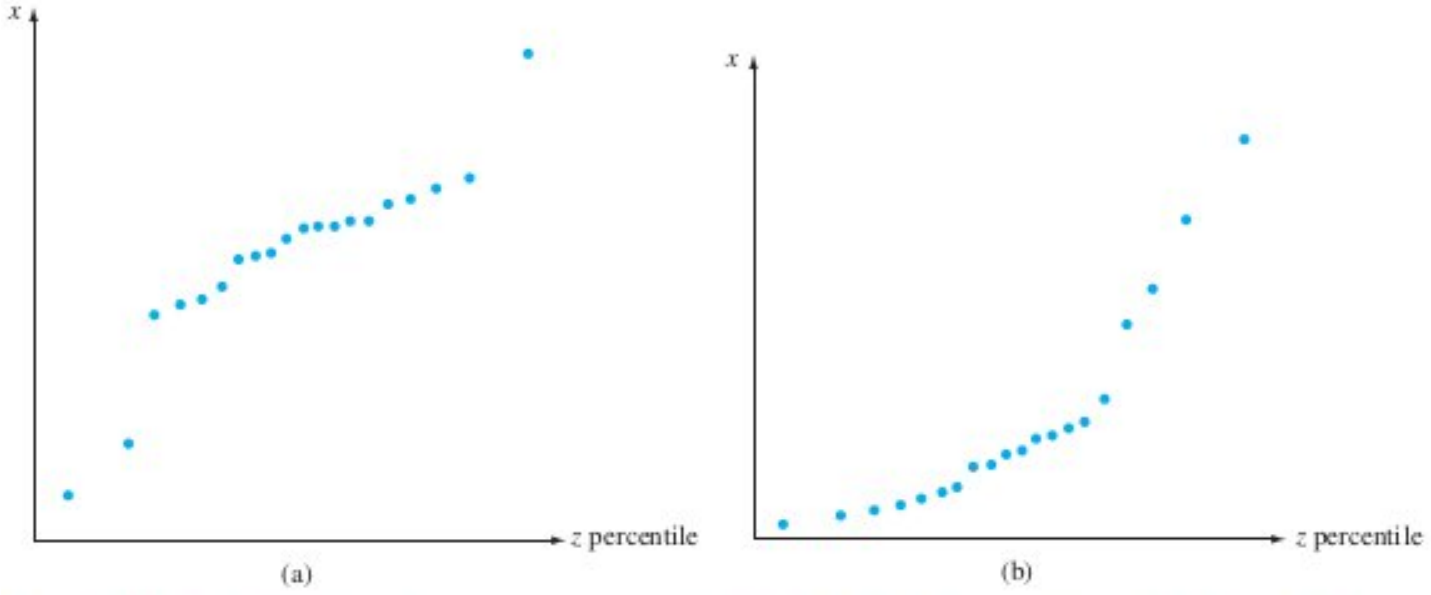


Figure 4.37 Probability plots that suggest a nonnormal distribution: (a) a plot consistent with a heavy-tailed distribution; (b) a plot consistent with a positively skewed distribution