4. The normal approximation to discrete distributions, and assessing normality

Readings:	Rosner: 5.6-5.9 OpenIntro Statistics: 3.1-3.2 Chihara and Hesterberg: 4.3
R:	stem, boxplot, qqnorm, qqline, shapiro.test, lillie.test, cvm.test, ad.test
SAS:	PROC UNIVARIATE
Homework:	Homework 2 due by 11:59 pm on September 17 Homework 3 due by 11:59 pm on September 24

Overview

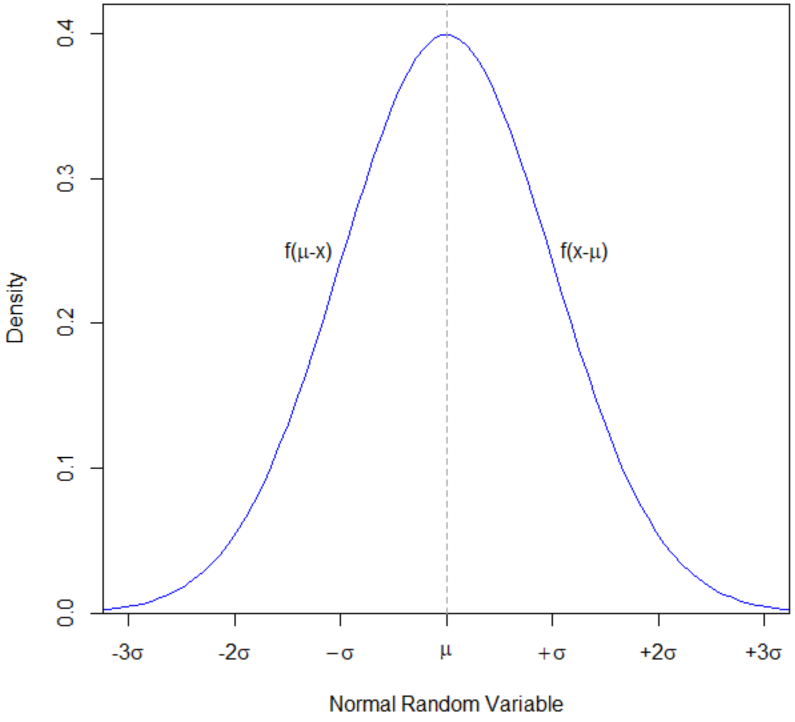
A) Normal distribution and properties
B) Normal approximation to the binomial and Poisson
C) Assessing normality
D) Other continuous distributions we will encounter

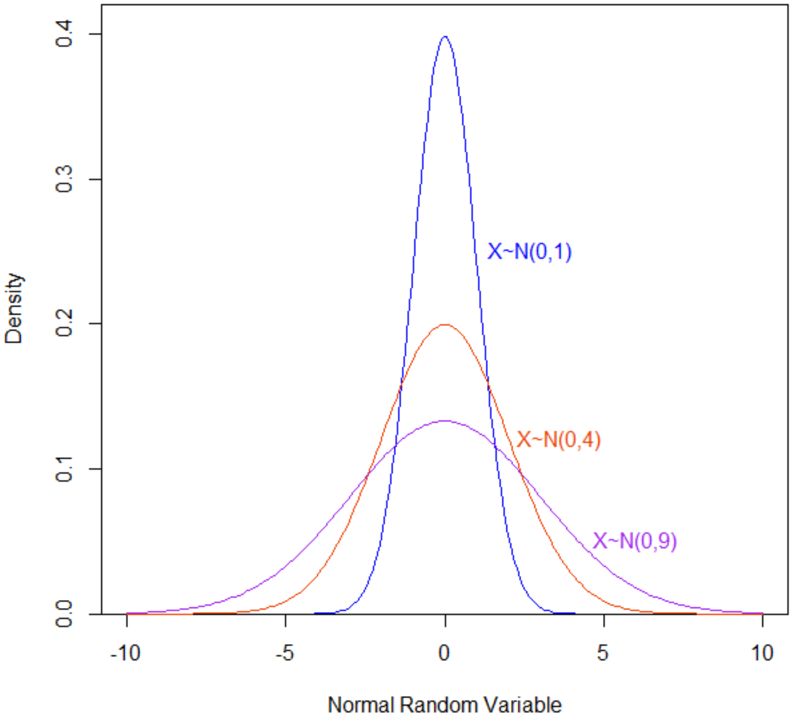
A) Normal distribution properties

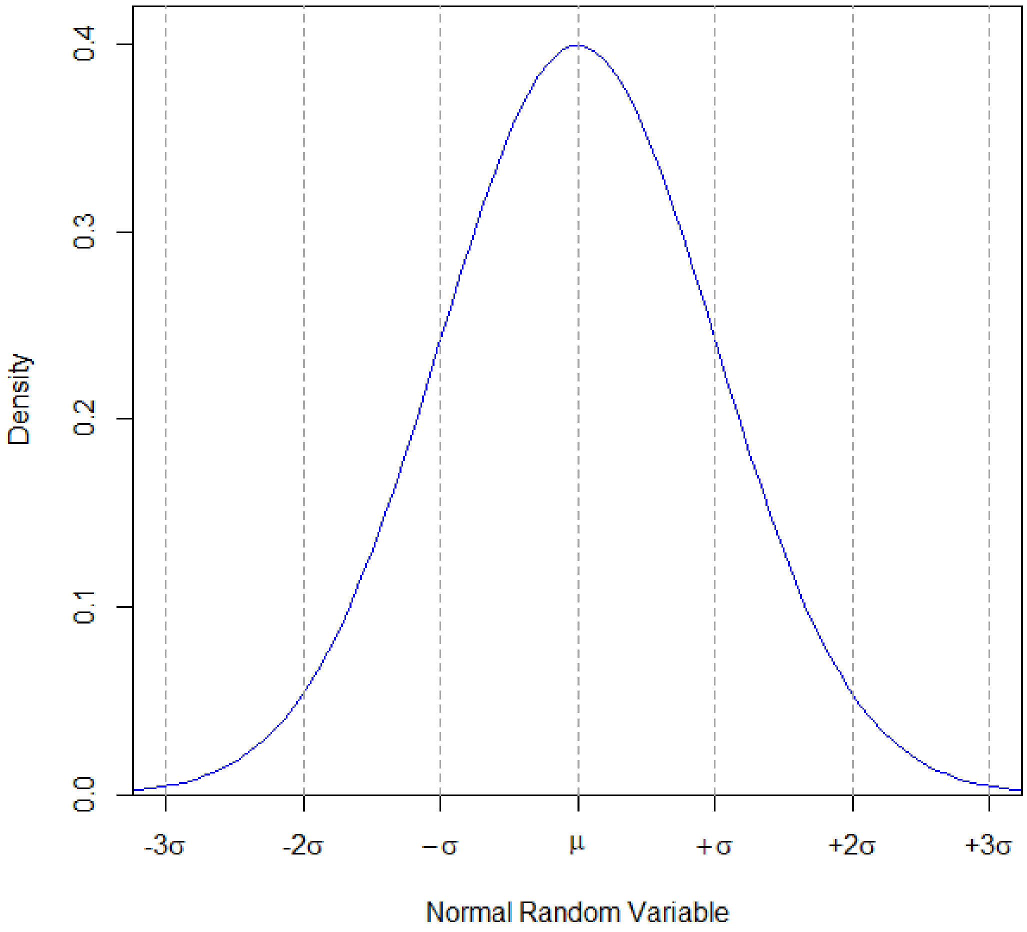
The normal distribution is defined by its *probability density function* (pdf), which is given as:

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

for parameters μ (mean) and $\sigma^2>0$ (variance). Its distribution is symmetric about μ: $f(\mu+x) = f(\mu-x)$.







P(μ - σ < X < μ + σ) = 0.6827 = (about 68% of area lies between ± 1 s.d.)

P(μ - 2σ < X < μ + 2σ) = 0.9545 (about 95% of area lies between ± 2 s.d.)

P(μ - 3σ < X < μ + 3σ) = 0.9973 (about 99.7% of area lies between ± 3 s.d.)

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.

$E(X) = \mu$ and $E(X^2)$ are population moments about zero

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

From the central moments, we can calculate standardized moments by dividing the k^th standard deviation (with various iterations for different approaches):

Skewness (3rd central moment): $\frac{E\left[(X-\mu)^3\right]}{s^3}$

Kurtosis (4th central moment): $\frac{E[(X-\mu)^4]}{s^4}$

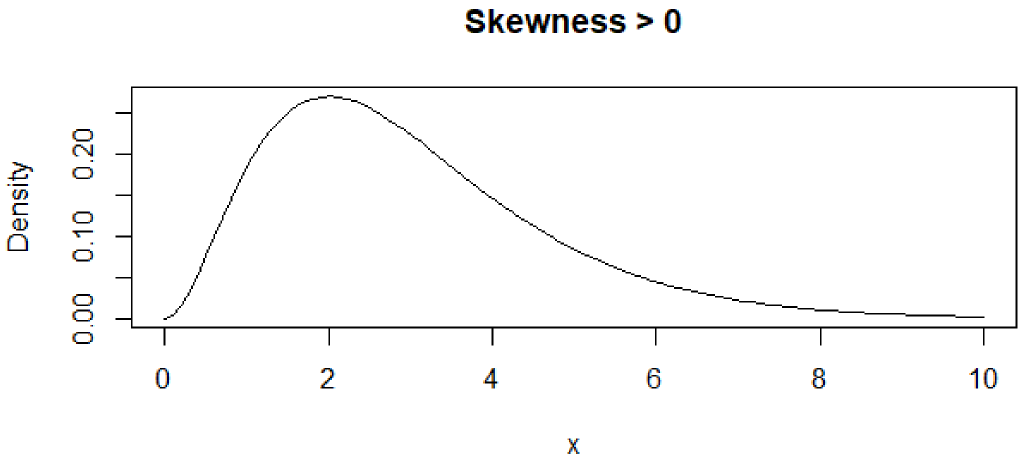
Skewness describes the symmetry of distribution.

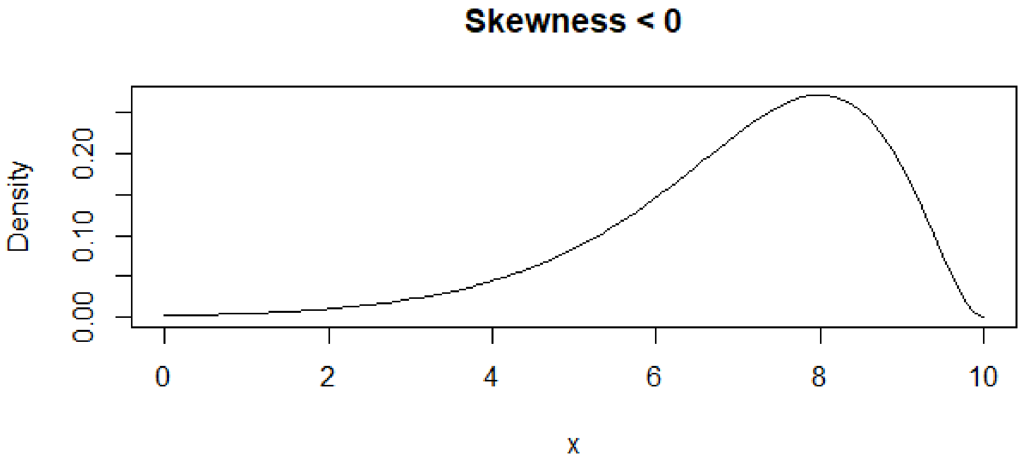
The 3rd central moment of the data, as with the 1st central moment, will balance out from left to right if the data are symmetric.

With normally distributed data, we expect skewness to be 0 (i.e., balanced).

If skewness is > 0: positive skew;
skewed to the right; more common

If skewness if < 0: negative skew;
skewed to the left



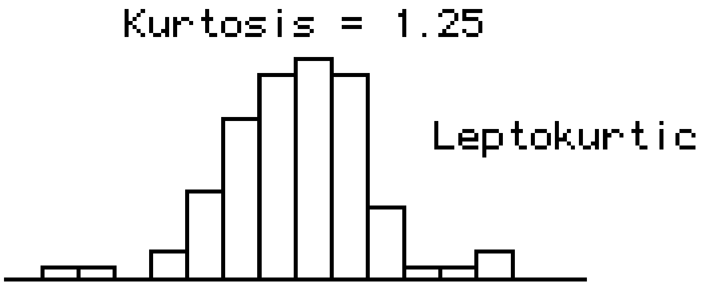


Kurtosis describes the "tailedness" of the probability distribution. Often we are interested in the *excess kurtosis*, which is Kurt[X] – 3

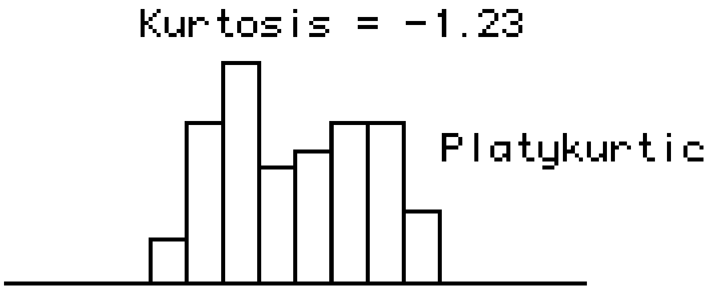
Excess kurtosis = 0, tails just like a normal distribution (*mesokurtic*), such as the (unsurprisingly) normal distribution and binomial when $p = \frac{1}{2} \pm \sqrt{\frac{1}{12}}$

Excess kurtosis > 0 heavier/fatter tails than a normal distribution (*leptokurtic*), such as the Student's t, exponential, and Poisson distributions

Excess kurtosis < 0 lighter/thinner tails than a normal distribution (*platykurtic*), such as the continuous & discrete uniform and Bernoulli distributions



> 0, heavier tails



< 0, thinner tails

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

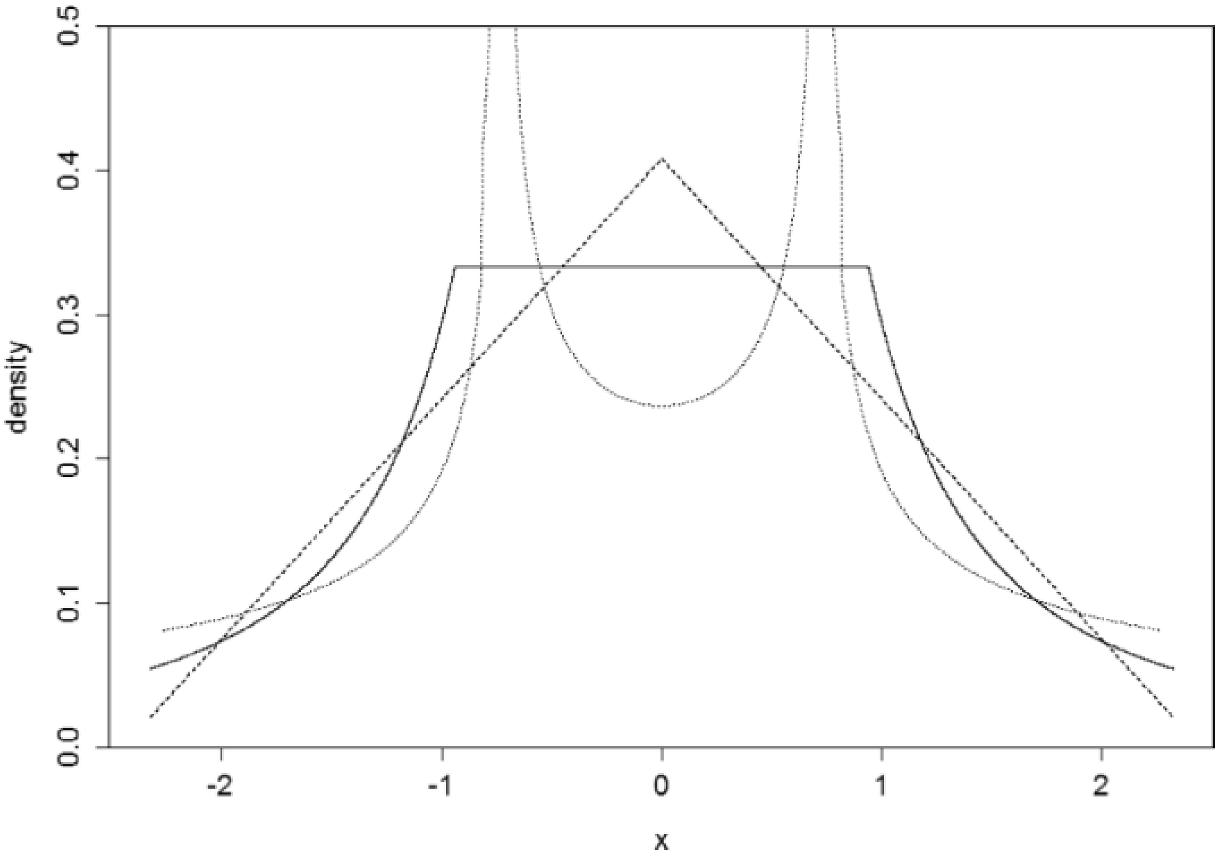


Figure 2.
Distributions with identical kurtosis = 2.4: solid = devil's tower, dashed = triangular, dotted = slip-dress.

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

B) Normal Approximation to Binomial and Poisson Distributions

Let's now see how the normal distribution is a limiting distribution for (or approximation to) the discrete distributions that we've talked about, the binomial and Poisson. This helps us to understand the how these distributions are similar and how they are different.

We know that the Normal distribution is a symmetric distribution with certain probabilities in specific parts of the distribution. Because of this, the normal will only approximate binomial and Poisson distributions that are, as you would suspect, fairly *symmetric*.

Recall that the binomial distribution has parameters *n* and *p* that completely define it, and that $\mu = np$, and $\sigma^2 = np(1-p)$. We know the variance of the binomial is maximized at p = 0.5. Binomial distributions with p close to 0.5 are the closest to symmetry.

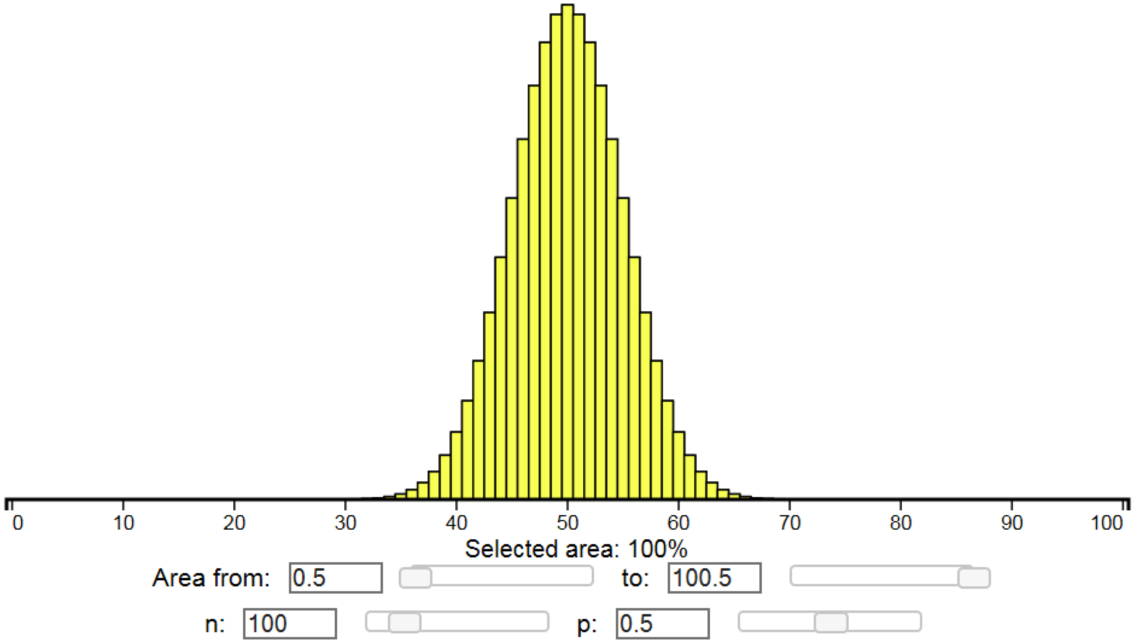
In general, if $np(1-p) \geq 5$, the binomial distribution looks fairly symmetric and is a possible candidate for a normal approximation.

Since the binomial is discrete, but the normal is continuous, how do we spread out the probability mass in the binomial over an infinite number of points? We will apply a *correction for continuity*.

Here's how the continuity correction works:

In the binomial, we can find P(X = x), but in the normal we know that this must somehow become an interval of values P(a < X < b). The interval that's used in practice is:
P(x-0.5 < X < x+0.5), we take a half step up from *x* and a half step down from *x* to capture the probability area in the normal distribution that belongs to the value *x* in the binomial distribution. In this way we can capture every bit of probability under the curve.

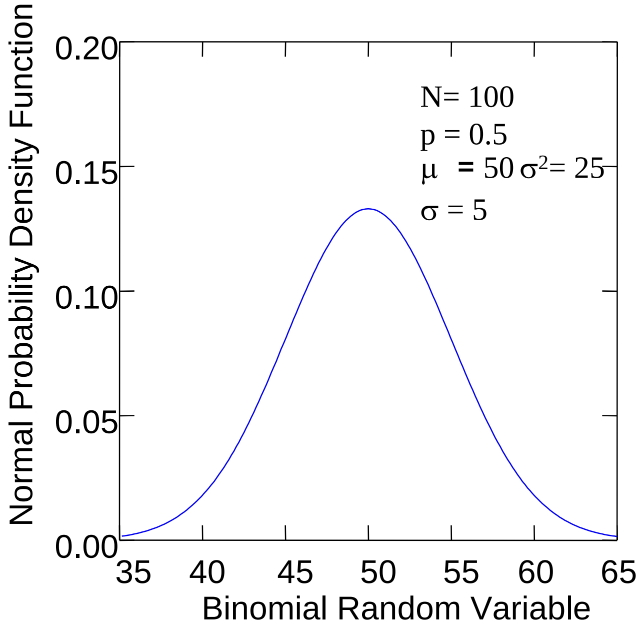
https://www.stat.berkeley.edu/~stark/Java/Html/BinHist.htm



Selected area: 100%

Area from: 0.5 to: 100.5

n: 100 p: 0.5



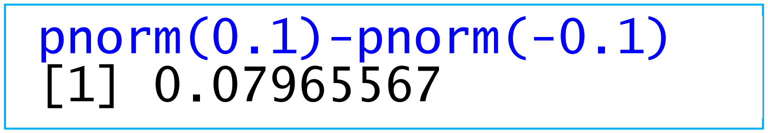
Now, if we want to find probabilities associated with the binomial, and the normal approximation conditions are met, we can use the Z-transformation, $Z = \frac{X-\mu}{\sigma}$, where Z ~ Normal(0, 1).

e.g. Consider a sample of 100 adult where the probability of having 1+ colds in a year is 0.5. What is the probability that 50 adults have 1+ cold in a year?

We know that *np* = μ = 100*0.50 = 50 and $\sigma = \sqrt{np(1-p)} = \sqrt{100 \times 0.5 \times 0.5} = 5$

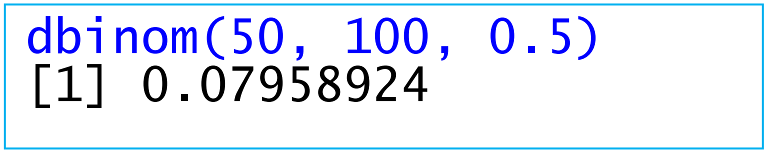
Applying the continuity correction, our probability becomes

$$\begin{aligned} P(50 - 0.5 \leq X \leq 50 + 0.5) \quad &= P(49.5 \leq X \leq 50.5) \\ &= P\left(\frac{49.5 - 50}{5} \leq Z \leq \frac{50.5 - 50}{5}\right) \\ &= P(-0.1 \leq Z \leq 0.1) \\ &= 0.07965567 \end{aligned}$$



```
pnorm(0.1)-pnorm(-0.1)
[1] 0.07965567
```

How does this compare to its exact Binomial probability?



```
dbinom(50, 100, 0.5)
[1] 0.07958924
```

(note that accuracy is within +/- 0.001)

What if we <u>didn't</u> apply the correction for continuity? $P(X=50) =$ 0

What about $P(X \geq 55)$?

This becomes $P(X \geq 54.5) = P(Z \geq (54.5-50)/5) = P(Z \geq 0.9) = 0.184061$
How does this compare to its exact Binomial probabilities?

```
1 - pnorm(0.9)
[1] 0.1840601

sum(dbinom(55:100, 100, 0.5))
[1] 0.1841008
```

What if we <u>didn't</u> apply the correction for continuity?
$P(X > 55) = P(Z > (55-50)/5) = P(Z > 1) = 0.1586553$

```
1 - pnorm(1.0)
[1] 0.1586553
```

What about $P(X \leq 40)$? (obtain on your own …)

P(X<=40)=0.0287

Normal Approximation to the Poisson

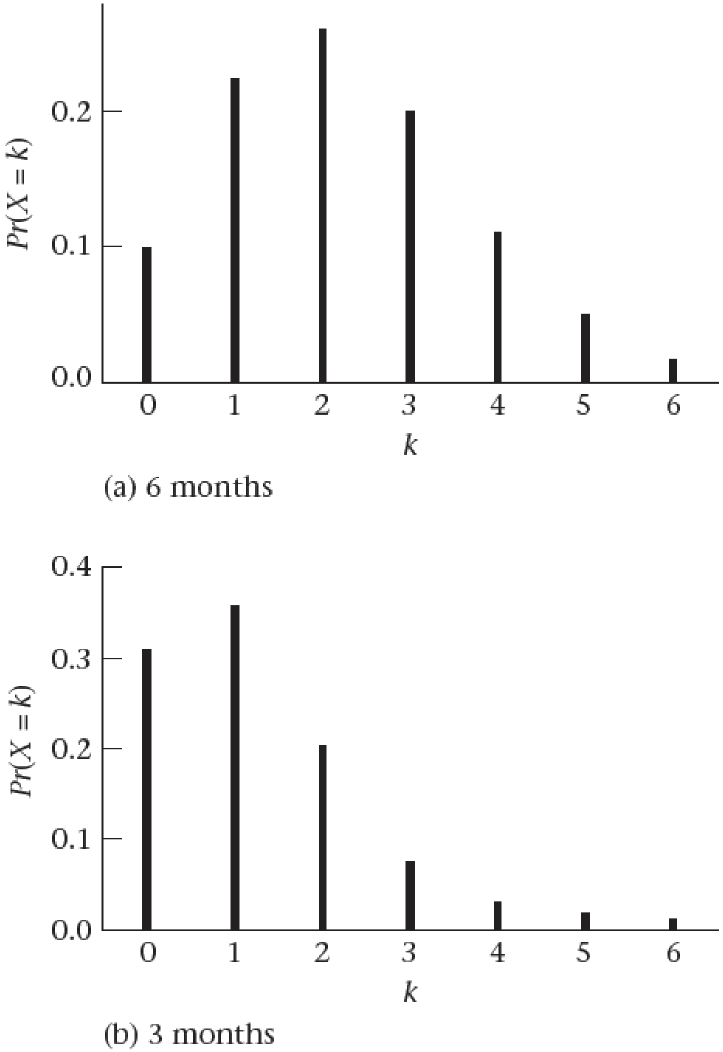
Now, let's see what conditions must hold for the normal distribution to approximate the Poisson.

Recall that the Poisson distribution is defined by one parameter, λ, and that $\mu = \sigma^2 = \lambda$.

We know that Poisson distributions can be very right-skewed, with lots of probability mass on low values and little probability mass on large values.

So, it will take a large value of λ to produce a symmetric distribution. λ ≥ 10 *generally* achieves this.

Figure 4.5 Distribution of the number of deaths attributable to typhoid fever over various time intervals

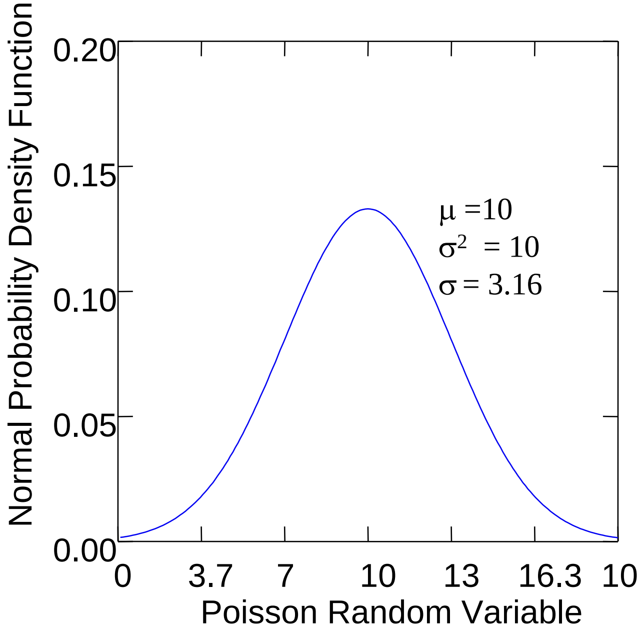


(a) 6 months

(b) 3 months

Rosner, Fundamentals of Biostatistics, 7th edition, 2010

We apply the same kind of continuity correction for finding Poisson probabilities based on the normal approximation as we do for the binomial.



Let’s say we want to find P(X = 10) – where X is the number of bacteria colonies growing on a culture dish. Assume λ = 10. Using the continuity correction we have

P(____ ≤ X ≤ ____) =

What is the exact Poisson probability for this example, and how does it compare to the approximation? *(Is the accuracy within +/- 0.001?)*

```
pnorm(0.16)-pnorm(-0.16)
[1] 0.1271189

dpois(10,10)
[1] 0.12511
```

What about P(X $\geq$ 15)?

With the continuity correction, P(X $\geq$ 14.5) = P(Z > 1.42________) = P(Z > _____) = 0.077

What is the exact Poisson probability for this example, and how does it compare to the approximation? *(Is the accuracy within +/- 0.001?)*

```
1 - pnorm(_____________)
[1] 0.07736446

1 - ppois(14, 10)
[1] 0.08345847
```

If λ = 1000, is the normal approximation's accuracy within +/- 0.001 for P(975 $\leq$ X $\leq$ 1050)?

always do 0.5 for the approximation.
poisson does not always get the potential approximation

```
pnorm(_______)-pnorm(_______)
[1] 0.7348499

ppois(1050,1000) - ppois(974,1000)
[1] 0.7334318
```

C) Assessing normality in a sample

Things to look for regarding normality of an empirical (sample) distribution:

Does the *mean = median = mode*? This will be true for any symmetric distribution. Is that enough to be called normal? ... no!

Other measures have to do with how skewed the distribution is. As we've talked about previously, a normal distribution has a *skewness* value of 0. A skewness value greater than 0 indicates a distribution is skewed to the right (+), less than 0 means skewed to the left (-).

Normal distributions have only so much probability mass in the middle and only so much out in the extremes or tails of the distribution. The measure called *kurtosis* measures (mostly, according to Peter Westfall's work) how light or heavy tailed a distribution is. A true normal distribution has a kurtosis value of 3 (i.e., an *excess kurtosis* of 0).

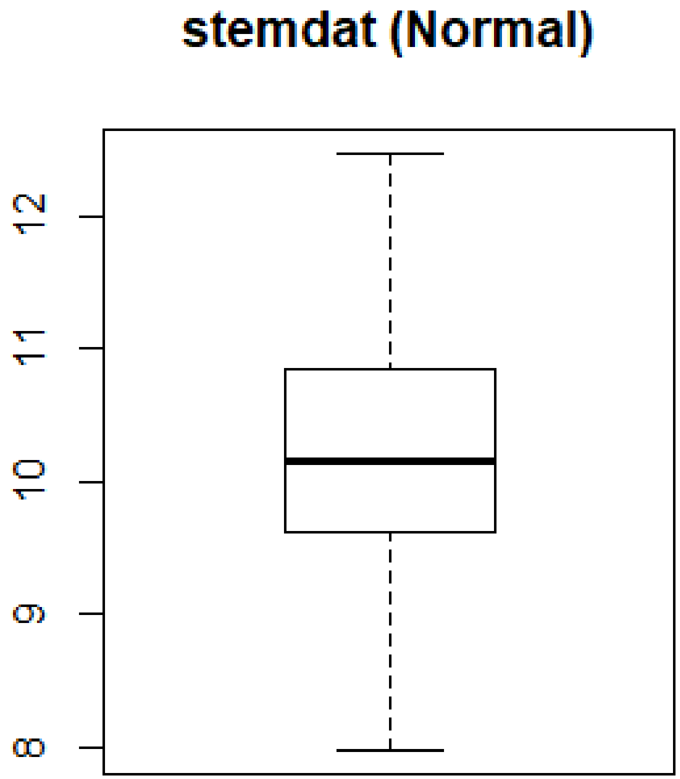
Plots are probably the best measures of normality - our eyes are really pretty good at pattern recognition. Be sure to use yours every time you approach a set of data!

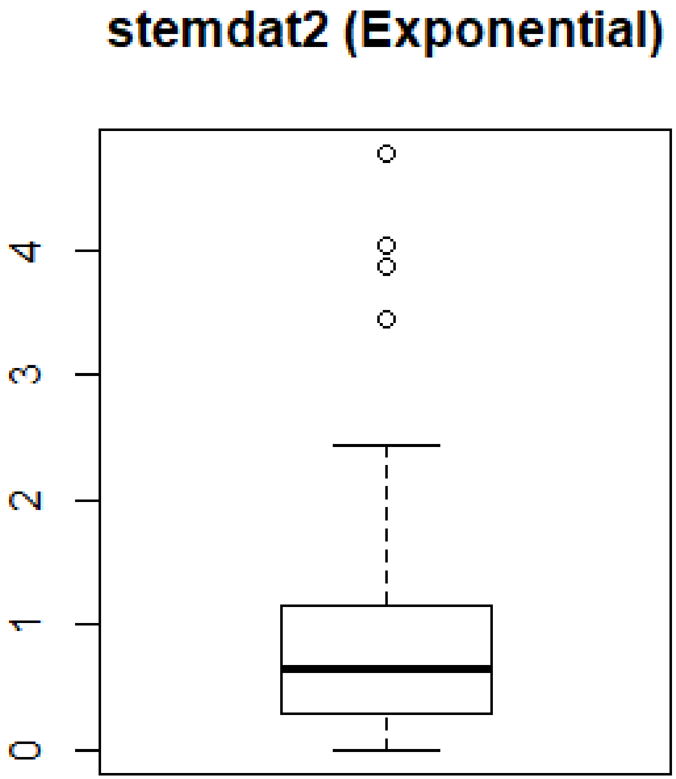
A ***Stem-and-Leaf*** plot looks like a histogram turned on its sides - when we smooth the histogram with our eyes do we see symmetry? Do we see a too heavy-tailed or a too light-tailed distribution?

```
set.seed(515)

stemdat <- rnorm(50, mean=10)

stem(stemdat)

  The decimal point is at the |

   7 |
   8 | 002
   8 | 669
   9 | 1123
   9 | 566777888889
  10 | 00123334
  10 | 5555677899
  11 | 01134
  11 | 58
  12 | 12
  12 | 5
```

```
set.seed(515)

stemdat2 <- rexp(50, rate=1)

stem(stemdat2)

  The decimal point is at the |

  0 | 0001112222233344444
  0 | 555666777777888
  1 | 012223
  1 | 5799
  2 | 04
  2 |
  3 | 4
  3 | 9
  4 | 0
  4 | 8
```

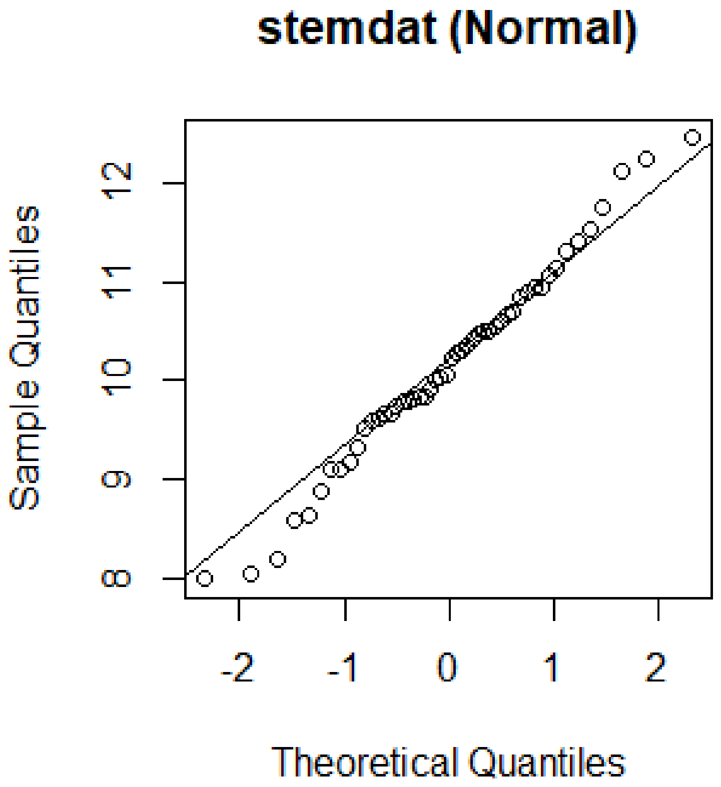
The *box plot* is also useful to us - the distances of the hinges to the median should be the same for a symmetric distribution. (R function: `boxplot`.)

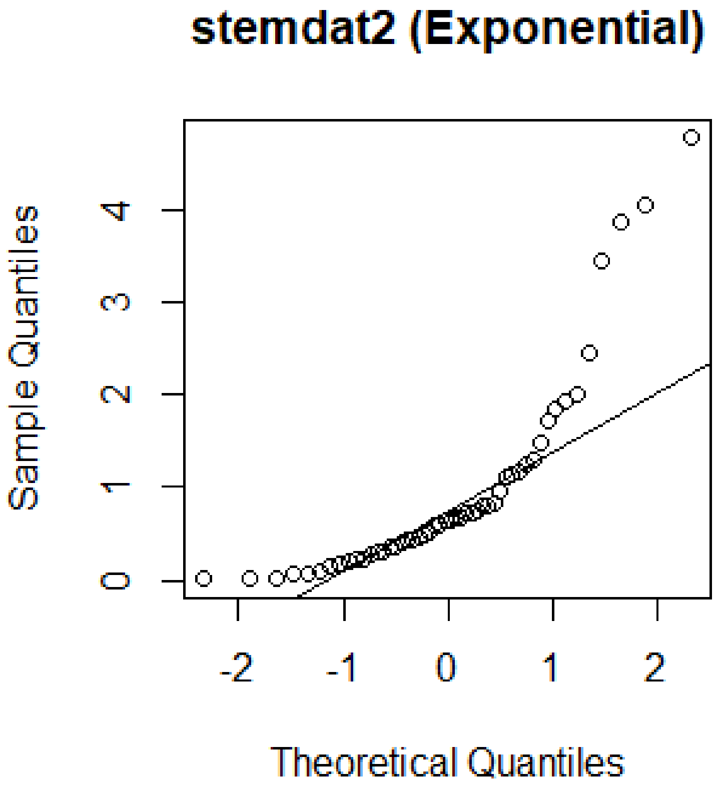




Another plot that we can look at is the *normal probability plot*. This is also known as a Q-Q plot and its shows actual (approximate) percentiles from the data versus expected percentiles (expressed as standard deviation units) of a normal distribution.

The expected percentiles are based on those of a normal distribution with $\mu = \bar{X}$ and $\sigma^2 = s^2$. Deviations from a straight line are informative. (R functions: `qqnorm` and `qqline`.)





Finally, there are a number of *statistical tests* that look for deviations from normality. Attached to these tests is a probability that you'd see a value of that statistic or one bigger (in most cases; or one smaller for the Shapiro Wilk test**) if the data are really, really normal, i.e. a *p-value* of this test.

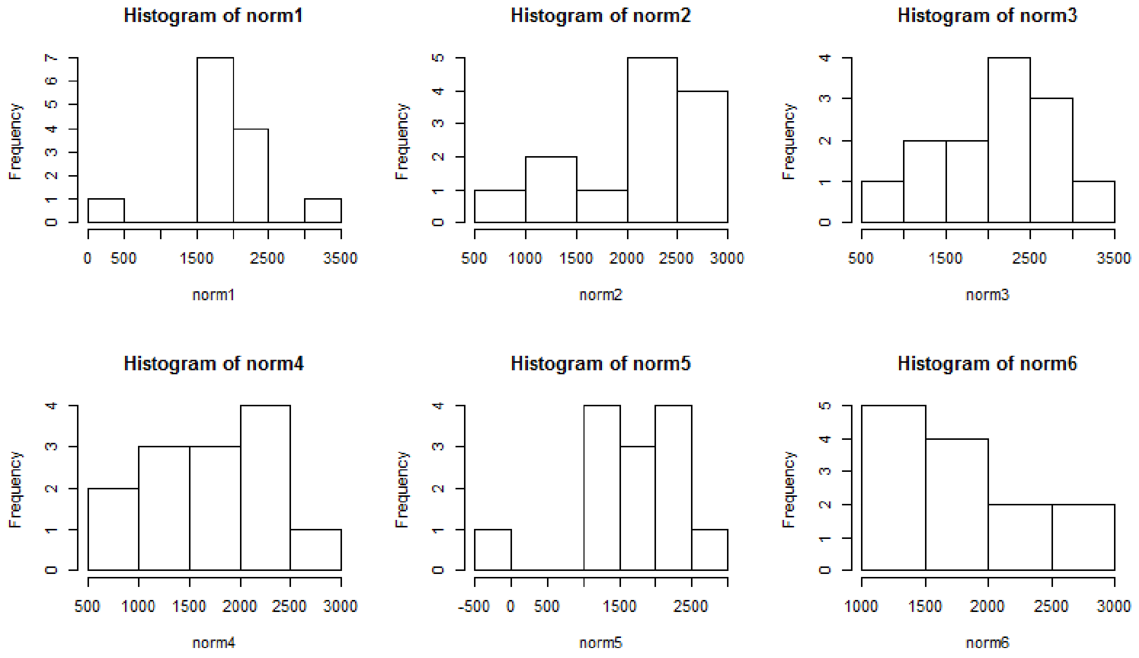
It's best *not* to weight the results of these tests too heavily - small sample sizes can lead to a decision of normality even when the data *are not* normal and large sample sizes can lead to a decision of non-normality even when the data *are* very close to normal.

Normality Test (Statistic)	**Function in R (Package)**	**stemdat Statistic; p-value**	**stemdat2 Statistic/p-value**
Shapiro-Wilk (W)**	`shapiro.test` (stats package)	W=0.987; p=0.846	W=0.753; p<0.001
Anderson-Darling (D)	`ad.test` (nortest package)	D=0.073; p=0.730	D=0.230; p<0.001
Cramer-von Mises (W)	`cvm.test` (nortest package)	W=0.028; p=0.871	W=0.711; p<0.001
Lilliefors/Kolmogorov-Smirnov (A)	`lillie.test` (nortest package)	A=0.192; p=0.892	A=4.087; p<0.001

Notice the difference in statistic values between our normal and exponential data sets, and how Shapiro-Wilk is significant for smaller W statistics.

One of the best ways to assess normality is through simulation. Generating a series of histograms from simulated data with a given mean and variance can remind us of the variety of patterns that are possible even when the underlying distribution is truly normal.

Based on a data set of dieters, if $\bar{x}$ = 1844 kcal and s = 638 kcal, here are six samples of size 13 randomly generated from a normal distribution with μ = 1844, σ = 638.



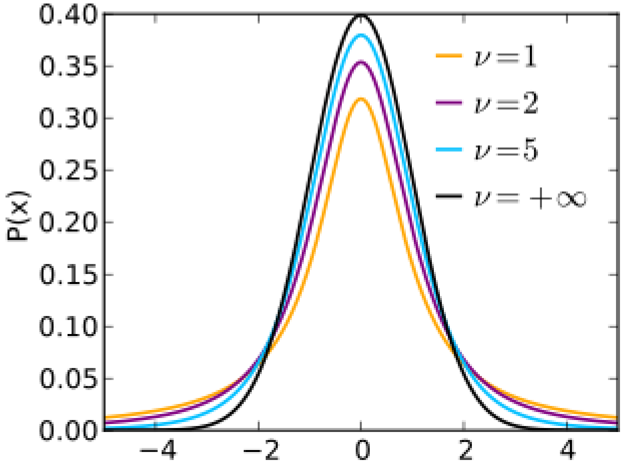
```
# Set up for 2x3 matrix of histograms
par(mfrow=c(2,3))

# Generate six random samples, histograms
set.seed(125)

norm1 <- rnorm(13, mean = 1844, sd = 638)
hist(norm1)

norm2 <- rnorm(13, mean = 1844, sd = 638)
hist(norm2)

norm3 <- rnorm(13, mean = 1844, sd = 638)
hist(norm3)

norm4 <- rnorm(13, mean = 1844, sd = 638)
hist(norm4)

norm5 <- rnorm(13, mean = 1844, sd = 638)
hist(norm5)

norm6 <- rnorm(13, mean = 1844, sd = 638)
hist(norm6)
```

To summarize, it’s best to take a *holistic* approach to assessing normality. Look at all of the available evidence and gather a full impression.

So why the heavy emphasis on the normal distribution? Along with the sentiments expressed by Youden, many of the statistical tests used in inferential statistics have their basis in the normal distribution.

THE
NORMAL
LAW OF ERROR
STANDS OUT IN THE
EXPERIENCE OF MANKIND
AS ONE OF THE BROADEST
GENERALIZATIONS OF NATURAL
PHILOSOPHY - IT SERVES AS THE
GUIDING INSTRUMENT IN RESEARCHES
IN THE PHYSICAL AND SOCIAL SCIENCES AND
IN MEDICINE AGRICULTURE AND ENGINEERING.
IT IS AN INDISPENSABLE TOOL FOR THE ANALYSIS AND THE
INTERPRETATION OF THE DATA OBTAINED BY OBSERVATION AND EXPERIMENT
W.J. YOUDEN

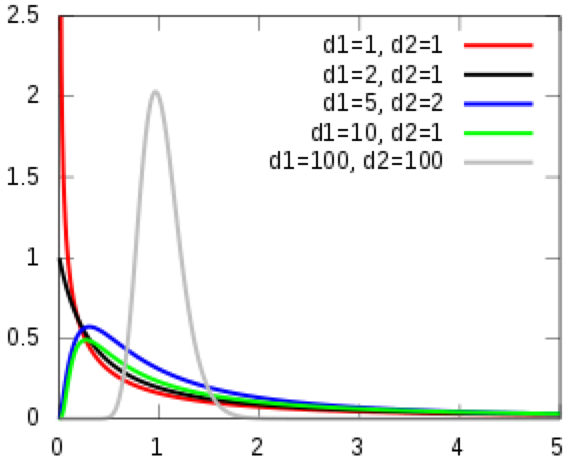
D) Other continuous distributions we will encounter



https://en.wikipedia.org/wiki/Student%27s_t-distribution

Student's *t*-distribution

- Symmetric and bell-shaped, but with fatter tails than the normal distribution
- Involves "degrees of freedom" (df=ν) that impact the heaviness of the tails (as $\nu \to \infty$, the t-distribution converges in distribution to the normal distribution)
- Used for the t-test, will appear during linear regression as the test statistic used to evaluate significance of the model coefficients



https://en.wikipedia.org/wiki/F-distribution

F-distribution

- If $X \sim t(\nu)$, then $X^2 \sim F(\nu_1 = 1, \nu_2 = \nu)$
- Has two parameters to describe its behavior
- Arises frequently in analysis of variance (ANOVA) (e.g., the *F*-test) and for comparing groups of predictors in a regression model simultaneously