Lecture 22: The Normal distribution: Kelly 6.2

(There is no lecture 21; Nov 18 was Midterm-2)

22.1 Standardizing random variables

Recall $\mathrm{E}(aY+b) \;=\; a\mathrm{E}(Y)+b$, $\mathrm{var}(aY+b) = a^2\mathrm{var}(Y)$.

Suppose Y is a discrete or continuous random variable with $\mathrm{E}(Y)=\mu$ and $\mathrm{var}(Y)=\sigma^2$. Let $Z=(Y-\mu)/\sigma$. Then $\mathrm{E}(Z) \;=\; (\mathrm{E}(Y)-\mu)/\sigma \;=\; (\mu-\mu)/\sigma \;=\; 0$, and $\mathrm{var}(Z) \;=\; (1/\sigma)^2\mathrm{var}(Y) \;=\; \sigma^2/\sigma^2 \;=\; 1$.

22.2 Location and scale parameters

A location parameter a shifts a probability density: the pdf is a function of $(x-a)$. For example, we can shift a uniform $U(0,1)$ pdf to a uniform $U(a,a+1)$ pdf.

A scale parameter stretches (or shrinks) a probability density. For example, to transform a Uniform $U(0,1)$ density to a Uniform $U(a,b)$, we shift by a and scale by $(b-a)$.

Another example of a scale parameter is the rate parameter of an exponential random variable. If we measure waiting times in minutes instead of hours, we still have an exponential shape to the pdf, but the rate per minute is 60 times less than the rate per hour.

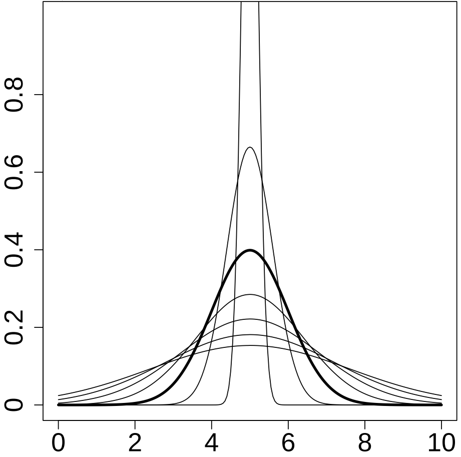
22.3 The Normal probability density, parameters μ and σ^2: $N(\mu,\sigma^2)$

$$f_X(x) \;=\; \frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2}) \qquad -\infty < x < \infty$$

$$P(X\in B_x) \;=\; \int_{B_x}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{(x-\mu)^2}{2\sigma^2})\;dx$$

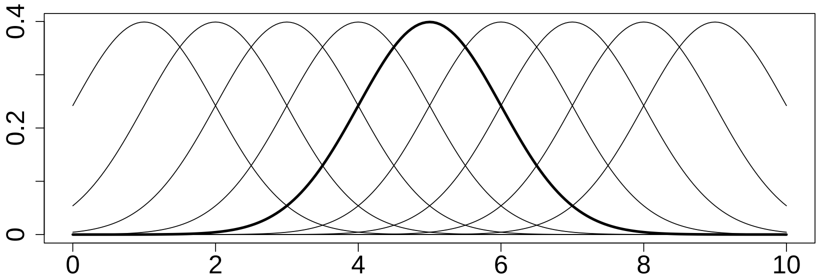
Let $z \;=\; (x-\mu)/\sigma$, $dz = dx/\sigma$, $B_z=\{(x-\mu)/\sigma; x\in B_x\}$

$$\begin{aligned} P(Z=(X-\mu)/\sigma\in B_z) &= \int_{B_z}\frac{1}{\sqrt{2\pi\sigma^2}}\exp(-\frac{z^2}{2})\;\sigma\;dz \\ &= \int_{B_z}\frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})\;dz \end{aligned}$$



That is $f_Z(z)$ is a Normal probability density with parameters 0 and 1.

Also, μ is a location parameter, σ is a scale parameter.



Above: Normal pdfs at different scales (values of σ).

Left: Normal pdfs at different locations (values of μ).

22.4 The standard Normal probability density

A random variable Z with p.d.f $f_Z(z) \;=\; \frac{1}{\sqrt{2\pi}}\exp(-\frac{z^2}{2})$ on $-\infty < z < \infty$ is a *standard Normal random variable.*

$\mathrm{E}(Z) \;=\; \int_{-\infty}^{\infty} z f_Z(z)dz \;=\; 0$ since $f_Z(-z)=f_Z(z)$.

So $\mathrm{var}(Z) \;=\; \mathrm{E}(Z^2) \;=\; \int_{-\infty}^{\infty} z^2 f_Z(z)dz$. In fact $\mathrm{E}(Z^2) \;=\; 1$ (not proved).

Now $Z=(X-\mu)/\sigma$ or $X=\mu+\sigma Z$, so for the general Normal $N(\mu,\sigma^2)$ random variable with parameters μ and σ^2, $\mathrm{E}(X)=\mu$ and $\mathrm{var}(X)=\sigma^2$.

Lecture 23: Normal approximation to the Binomial; Kelly 6.4

23.1: Stirling's formula

For large n, $n!$ is approximately $n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}$. Let $k = np$, then

$$\begin{aligned}
\binom{n}{k} &= \frac{n!}{k!(n-k)!} = \frac{n!}{(np)!(n(1-p))!} \\
&\approx \frac{n^{n+\frac{1}{2}}e^{-n}\sqrt{2\pi}}{(np)^{np+\frac{1}{2}}e^{-np}\sqrt{2\pi}(n(1-p))^{n(1-p)+\frac{1}{2}}e^{-n(1-p)}\sqrt{2\pi}} \\
&= \frac{n^{n+\frac{1}{2}}}{(np)^{np+\frac{1}{2}}(n(1-p))^{n(1-p)+\frac{1}{2}}\sqrt{2\pi}} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{np(1-p)}}\frac{1}{p^{np}(1-p)^{n(1-p)}}
\end{aligned}$$

23.2: Standardizing the binomial probabilities

Suppose X has the Binomial $Bin(n,p)$ p.m.f.

Then $\mathrm{E}(X) = np$, $\mathrm{var}(X) = np(1-p)$.

So if $Z = (X-np)/(np(1-p))$, then $\mathrm{E}(Z) = 0$, $\mathrm{var}(Z) = 1$.

23.3: The DeMoivre-Laplace limit theorem

(a) For a $Bin(n,p)$ random variable X, the p.m.f. is largest at $k \approx np$: $P(X = k) = \binom{n}{k}p^k(1-p)^{n-k}$.

So for n large and $k = np$ we have

$$P(X = np) = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{np(1-p)}}\frac{1}{p^{np}(1-p)^{n(1-p)}}\ p^{np}\ (1-p)^{n(1-p)} = \frac{1}{\sqrt{2\pi}}\frac{1}{\sqrt{np(1-p)}}$$

But this is also the value of the maximum pdf of a Normal random variable with mean $\mu = np$ and variance $\sigma^2 = np(1-p)$.

(b) As X increases from k to $k+1$, $Z = (X-np)/\sqrt{np(1-p)}$ increases from z to $z+\delta$ where $z = (k-np)/\sqrt{np(1-p)}$ and $\delta = 1/\sqrt{np(1-p)}$. (Note $np(1-p) = 1/\delta^2$.) For a Normal $N(np, 1/\delta^2)$ pdf;

$$\frac{f_Y(k+1)}{f_Y(k)} = \frac{\frac{\delta}{\sqrt{2\pi}}\exp(-\delta^2(k-np+1)^2/2)}{\frac{\delta}{\sqrt{2\pi}}\exp(-\delta^2(k-np)^2/2)} = \frac{\exp(-\delta^2((z/\delta)+1)^2/2)}{\exp(-\delta^2(z/\delta)^2/2)} = \exp(-z\delta - \delta^2/2) \approx (1-\delta z)$$

(c) But

$$\frac{P(X=(k+1))}{P(X=k)} = \frac{\binom{n}{k+1}p^{k+1}(1-p)^{n-k-1}}{\binom{n}{k}p^k(1-p)^{n-k}} = \left(\frac{n-k}{k+1}\right)\left(\frac{p}{1-p}\right)$$

$$\begin{aligned}
&= \left(\frac{n(1-p) - \ z/\delta}{np + (z/\delta) + 1}\right)\left(\frac{p}{1-p}\right) = \frac{np(1-p) - zp/\delta}{np(1-p) + (1-p)z/\delta\ +\ (1-p)} \approx \frac{1-\delta zp}{1+\delta z(1-p)} \\
&\approx (1-\delta zp)(1-\delta z(1-p)) = (1-\delta zp - \delta z(1-p)) = (1-\delta z)
\end{aligned}$$

23.4: A preview of the Central Limit Theorem

Recall that $Bin(n,p)$ is sum of n independent Bernoulli, each with mean p and variance $p(1-p)$.

Suppose, $Y_1, ..., Y_n$ are independent, with the same distribution, each with mean μ and variance σ^2, and $T_n = \sum_{i=1}^n Y_i$. Then $\mathrm{E}(T_n) = n\mu$ and $\mathrm{var}(T_n) = n\sigma^2$.

$T_n^* = (T_n - n\mu)/(\sqrt{n}\ \sigma)$ has mean 0 and variance 1.

Subject to some conditions, the same result holds for T_n as for the $Bin(n,p)$. That is T_n^* has approx. a $N(0,1)$ pdf. This is the **Central Limit Theorem**.

Lecture 24: Using the Normal approximation

24.1: Using the Normal probability table

The table in Kelly, Appendix C, P.601, is of the usual form for the probabilities for a N(0,1) standard Normal distribution. It gives $P(Z \leq x)$ for values of x from 0 up. This probability is denoted $\Phi(x)$.

For negative x, $P(Z \leq x) \ = \ P(Z \geq -x) \ = \ 1 - P(Z \leq -x)$.

(Note that since Z is a continuous random variable, $P(Z < x) \ = \ P(Z \leq x)$).

24.2: The continuity correction

(a) When approximating $P(X = k)$ for a binomial X by a Normal Y, we must consider

$P(k - \frac{1}{2} < Y \leq k + 1/2)$. We need area under the curve.

(b) Hence when we approximate $P(a \leq X \leq b)$ we should use $P(a - \frac{1}{2} < Y < b + \frac{1}{2})$.

(c) However, for large n and a reasonable range of Y it makes almost no difference. Recall when X is increased by 1, Z is increased by $\delta = 1/\sqrt{np(1-p)}$.

24.3 Approximating discrete Binomial with continuous Normal

(c) Example: suppose X is $Bin(30, 2/3)$. $\mathrm{E}(X) = 20$, $\mathrm{var}(X) = 30 \times (1/3) \times (2/3) \ = \ 20/3$.

Compute the probability $14 \leq X \leq 18$

(i) Exactly, using the Binomial probabilities: $\sum_{k=14}^{18} P(X = k)$. Answer 0.2689.

(ii) Using the Normal approx, with the range 14 to 18 for X:

$Z = (14 - 20)/\sqrt{20/3} = -2.32$ to $Z = (18 - 20)/\sqrt{20/3} = -0.77$. Answer: 0.2105.

(iii) Using the Normal approx, with the range 13.5 to 18.5 for X:

$Z = (13.5 - 20)/\sqrt{20/3} = -2.517$ to $Z = (18.5 - 20)/\sqrt{20/3} = -0.5809$. Answer: 0.2747.

For general a, b: $P(a < Z \leq b) \ = \ \Phi(b) - \Phi(a)$.

24.4: Mendel's experiments

Mendel did many experiments of the form of the one with the red/white flowers. He crossed red-flowered plants with white-flowered plants, so he knew the red-flowered offspring were of RW type. These are known as the F_1 or *hybrids*. He then crossed these with each other, and expected to get red and white flowers in the ratio 3:1. Here are four examples:

a) 253 F_1 producing 7324 seeds: 5474 round, 1850 wrinkled: ratio 2.96:1

b) 258 F_1 producing 8023 seeds: 6022 yellow, 2001 green: ratio 3.01:1.

c) 929 F_2; 705 red flowers, 224 white flowers: ratio 3.15:1.

d) 580 F_2: 428 green pods, 152 yellow pods: ratio 2.82:1

24.5 Are Mendel's results too good?

There has been much debate as to whether Mendel's results are "too good" – too close to the 3:1 ratio.

Note the larger samples for characteristics that can be observed at the seed stage. These give the ratios closest to 3:1. This is as expected: $\mathrm{var}(X) = np(1-p)$ but $\mathrm{var}(X/n) \ = \ \mathrm{var}(X)/n^2 = \ p(1-p)/n$ which decreases as n increases. Are we too close? Recall $Z \ = \ (X - np)/\sqrt{np(1-p)}$ is approx N(0,1). Here $p = 3/4$:

a) $Z_a \ = \ (5474 - 7324 \times 0.75)/\sqrt{7324 \times 3/16} \ = -0.5127$, $P(-0.5127 < Z \leq 0.5127) \ = 2\Phi(0.5127) - 1 = \ 0.39$.

b) $Z_b \ = \ (6022 - 8023 \times 0.75)/\sqrt{8023 \times 3/16} \ = \ 0.1225$, $P(-0.1225 < Z \leq 0.1225) \ = \ 2\Phi(0.1225) - 1 \ = \ 0.097$.

c) $Z_c \ = \ (705 - 929 \times 0.75)/\sqrt{929 \times 3/16} \ = \ 0.6251$, $P(-0.6251 < Z \leq 0.06251) \ = \ 2\Phi(0.6251) - 1 \ = \ 0.468$.

d) $Z_d = (428 - 580 \times 0.75)/\sqrt{580 \times 3/16} \ = \ -0.6712$, $P(-0.6712 < Z \leq 0.6712) \ = \ 2\Phi(0.6712) - 1 \ = \ 0.498$.

So far, with these experiments, there seems no reason to think Mendel's results are "too good".

Lecture 25: More examples from Mendel's experiments

25.1 Combining the experiments

The fact that these involve different characteristics does not stop us combining them. They are all independent Bernoulli trials with $p = 0.75$.

We have $7324 + 8023 + 929 + 580 \;=\; 16856$ trials with $5474 + 6022 + 705 + 428 \;=\; 12629$ "successes". $Z = (12629 - 16856 \times 0.75)/\sqrt{16856 * 3/16} \;=\; -0.2312$. $P(-0.2312 < Z \leq 0.2312) \;=\; 2\Phi(0.2312) - 1 \;=\; 0.183$.

Alternatively, we can combine the Z-values: we could do this even if they came from Bernoulli trials with different p. Here: $Z_a + Z_b + Z_c + Z_d = -0.5127 + 0.1225 + 0.6251 - 0.6712 \;=\; -0.4363$.

This would be a Normal with mean 0 but variance 4 (why?). So we must standardize it:
$Z^* = -0.4363/2 = -0.2182$, $P(-0.2182 < Z \leq 0.2182) \;=\; 2\Phi(0.2182) - 1 \;=\; 0.173$.

So again, either way, here there is no evidence of the results being "too good". However, when a large number of Mendel's other results are also grouped together, overall, they do look a bit "too good".

25.2 Mendel's experiment: continued

Now Mendel wanted to show not just the 3:1 red:white ratio, but also the 1:2:1 for $RR : RW : WW$. So he needed to find which of his red-flowered F_2 plants were RR and which were RW. To do this he *selfed* his red-flowered F_2 pea plants: that is, the parents were RR giving $RR \times RR$ or RW giving $RW \times RW$.

In order to tell whether the parent was RW, Mendel grew up 10 offspring, and if all were red he said the plant *bred true*. Note, under Mendel's hypothesis $P(RR \mid \text{red}) \;=\; 1/3$.

Mendel reported his result: from 600 F_2 he found 201 *bred true*. Assuming 1/3 should *breed true*, is this result too close to 1/3? Note if $p = 1/3$, $\mathrm{E}(X) = 200$, $\mathrm{var}(X) = 600 \times 1/3 \times 2/3 = 400/3$.

(i) Without the correction (considering $X = 199, 200, 201$) show the probability of being this close is about 6.5%. ($Z = \pm 0.08660$).

(ii) With the correction ($198.5 < X < 201.5$) show the probability of being this close is a bit over 10% ($Z = \pm 0.12990$).

(Here the continuity correction makes enough difference that is might affect our belief about whether Mendel's results are "too good").

25.3 Mendel's mistake:

Recall that each offspring of an $RW \times RW$ mating is white with probability 1/4.

(i) For each $RW \times RW$ mating, what is the probability Mendel mis-called it as $RR \times RR$?
Answer: $(3/4)^{10} = 0.0563$.

(ii) If the frequency of RR parents is 1/3 and RW is 2/3, what is the overall probability that all 10 offspring plants are red? Answer: $(1/3) + (2/3) \times 0.0563 \;=\; 0.371$.

25.4 Probability of being close to 0.371

So now the p of Mendel's Binomial should have been $p = 0.371$. $\mathrm{E}(X) = 222.6$, $\mathrm{var}(X) = 140.01$, st.dev $= 11.83$. Now we need the probability that Mendel's reported count of 201 would be *this far off*.

(i) With no correction: $X \leq 201$, $Z < -1.825$ or $Z > 1.825$. Answer: about 6.8%.

(ii) With correction: $X \leq 201.5$, $Z < -1.783$ or $Z > 1.783$. Answer: about 7.4%.

(iii) Or maybe we should ask, this far off in direction of his assumed 1/3, Asnwers: 3.4% and 3.7%.

Either Mendel was, for once, quite *unlucky* or else his result is too close to what he may have expected, and too far from what he should have found.