Lecture 4: Normal Distribution / Binomial Approximation

Sta230/Mth230

Colin Rundel

January 24, 2014

Taylor Series

For any* function $f(x)$ we can rewrite it using

$$f(x) = f(a) + \frac{f'(a)}{1!}(x-a) + \frac{f''(a)}{2!}(x-a)^2 + \frac{f'''(a)}{3!}(x-a)^3 + \cdots$$

Often each progressive term gets smaller, so we can approximate the function with only the first several terms

$$f(x) \approx f(a) + \frac{f'(a)}{1!}(x-a)$$

.

Example - $\log(1+x)$

Find the Taylor expansion of $f(x) = \log(1+x)$.

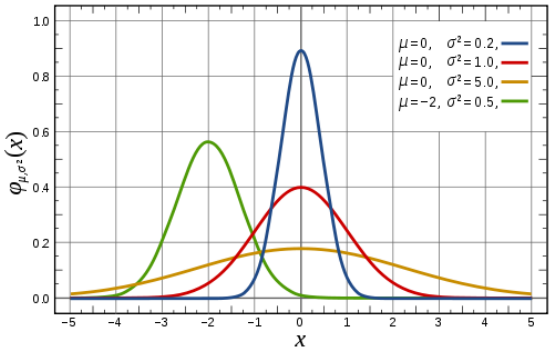
Therefore, when x is small

$$log(1+x) \approx x$$

Normal Distribution

If X is random variable with a normal distribution with a mean μ and variance σ^2, $X \sim \mathcal{N}(\mu, \sigma^2)$, then

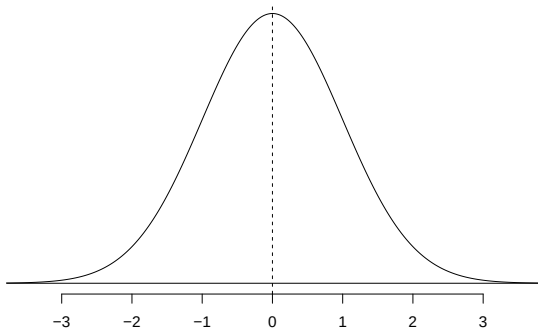
$$f(x|\mu,\sigma^2) = \frac{1}{\sqrt{2\pi}\sigma}\ e^{-\frac{1}{2}\frac{(x-\mu)^2}{\sigma^2}}$$



Unit Normal Distribution

The unit normal distribution is a special case of the normal distribution where $\mu = 0$ and $\sigma = 1$, $Z \sim \mathcal{N}(0, 1)$.

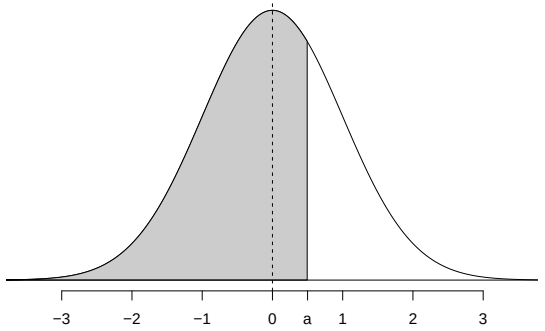
$$f(z) = \phi(z) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2}z^2}$$



Properties of the Unit Normal Distribution

The area under the unit normal curve from $-\infty$ to a is given by

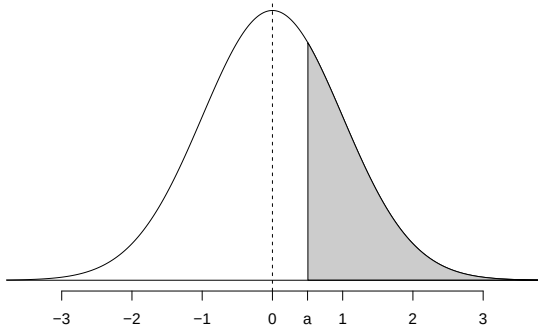
$$P(Z \leq a) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{a} e^{-t^2/2}\, dt = \Phi(a)$$



Properties of the Unit Normal Distribution

The area under the unit normal curve from a to ∞ is given by

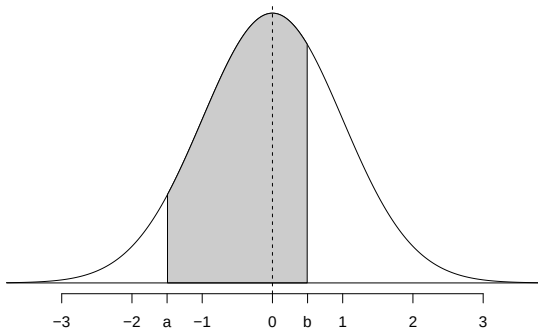
$$P(Z \geq a) = \frac{1}{\sqrt{2\pi}} \int_{a}^{\infty} e^{-t^2/2}\, dt = 1 - \Phi(a)$$



Properties of the Unit Normal Distribution

The area under the unit normal curve from a to b where $a \leq b$ is given by

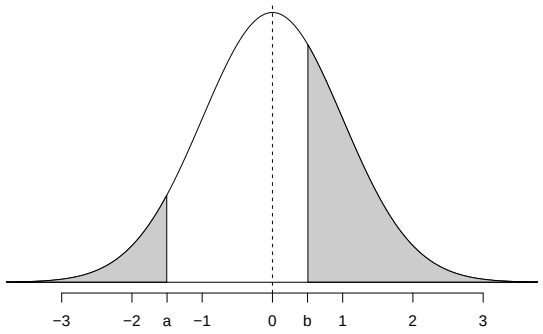
$$P(a \leq Z \leq b) = \frac{1}{\sqrt{2\pi}} \int_{a}^{b} e^{-t^2/2}\, dt = \Phi(b) - \Phi(a)$$



Properties of the Unit Normal Distribution

The area under the unit normal curve outside of a to b where $a \leq b$ is given by

$$P(a \geq Z \text{ or } Z \geq b) = 1 - \frac{1}{\sqrt{2\pi}} \int_a^b e^{-t^2/2}\, dt = \Phi(a) + (1 - \Phi(b)) = 1 - (\Phi(b) - \Phi(a))$$



Evaluating Φ

The function $\Phi(x)$ has no simple closed form solution, sometimes it is written in terms of the error function erf

$$\Phi(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} e^{-t^2/2}\, dt = \frac{1}{2}\left[1 + \text{erf}\left(\frac{x}{\sqrt{2}}\right)\right]$$

where

$$\text{erf}(x) = \frac{2}{\sqrt{\pi}} \int_0^x e^{-t^2}\, dt$$

but this doesn't get us very far, but we can take the Taylor expansion of e^{-t^2} and evaluate the integral of each term to get a series for erf(x).

Approximating erf(x)

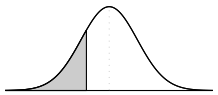
First we expand $f(x) = e^{-x^2}$,

Approximating $\Phi(x)$

```
erf = function(x) {
    (2/sqrt(pi)) * (x - x^3/3 + x^5/10 - x^7/42 + x^9/216)
}

Phi = function(x) {
    0.5 * (1 + erf(x/sqrt(2)))
}

Phi(1.5)

## [1] 0.9339

pnorm(1.5)

## [1] 0.9332
```

Evaluating Φ in practice

In practice you will never have to evaluate Φ explicitly, we either look $\Phi(x)$ up in a table or use a computer to calculate it (`pnorm` in `R`).



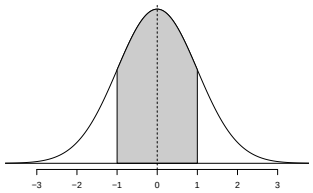
Second decimal place of Z										
0.09	0.08	0.07	0.06	0.05	0.04	0.03	0.02	0.01	0.00	Z
0.0002	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	−3.4
0.0003	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0005	0.0005	0.0005	−3.3
0.0005	0.0005	0.0005	0.0006	0.0006	0.0006	0.0006	0.0006	0.0007	0.0007	−3.2
0.0007	0.0007	0.0008	0.0008	0.0008	0.0008	0.0009	0.0009	0.0009	0.0010	−3.1
0.0010	0.0010	0.0011	0.0011	0.0011	0.0012	0.0012	0.0013	0.0013	0.0013	−3.0
0.0014	0.0014	0.0015	0.0015	0.0016	0.0016	0.0017	0.0018	0.0018	0.0019	−2.9
0.0019	0.0020	0.0021	0.0021	0.0022	0.0023	0.0023	0.0024	0.0025	0.0026	−2.8
0.0026	0.0027	0.0028	0.0029	0.0030	0.0031	0.0032	0.0033	0.0034	0.0035	−2.7
0.0036	0.0037	0.0038	0.0039	0.0040	0.0041	0.0043	0.0044	0.0045	0.0047	−2.6
0.0048	0.0049	0.0051	0.0052	0.0054	0.0055	0.0057	0.0059	0.0060	0.0062	−2.5
0.0064	0.0066	0.0068	0.0069	0.0071	0.0073	0.0075	0.0078	0.0080	0.0082	−2.4
0.0084	0.0087	0.0089	0.0091	0.0094	0.0096	0.0099	0.0102	0.0104	0.0107	−2.3
0.0110	0.0113	0.0116	0.0119	0.0122	0.0125	0.0129	0.0132	0.0136	0.0139	−2.2
0.0143	0.0146	0.0150	0.0154	0.0158	0.0162	0.0166	0.0170	0.0174	0.0179	−2.1
0.0183	0.0188	0.0192	0.0197	0.0202	0.0207	0.0212	0.0217	0.0222	0.0228	−2.0
0.0233	0.0239	0.0244	0.0250	0.0256	0.0262	0.0268	0.0274	0.0281	0.0287	−1.9
0.0294	0.0301	0.0307	0.0314	0.0322	0.0329	0.0336	0.0344	0.0351	0.0359	−1.8
0.0367	0.0375	0.0384	0.0392	0.0401	0.0409	0.0418	0.0427	0.0436	0.0446	−1.7
0.0455	0.0465	0.0475	0.0485	0.0495	0.0505	0.0516	0.0526	0.0537	0.0548	−1.6
0.0559	0.0571	0.0582	0.0594	0.0606	0.0618	0.0630	0.0643	0.0655	0.0668	−1.5
0.0985	0.1003	0.1020	0.1038	0.1056	0.1075	0.1093	0.1112	0.1131	0.1151	−1.2
0.1170	0.1190	0.1210	0.1230	0.1251	0.1271	0.1292	0.1314	0.1335	0.1357	−1.1
0.1379	0.1401	0.1423	0.1446	0.1469	0.1492	0.1515	0.1539	0.1562	0.1587	−1.0
0.1611	0.1635	0.1660	0.1685	0.1711	0.1736	0.1762	0.1788	0.1814	0.1841	−0.9

Evaluating Φ - Some practice

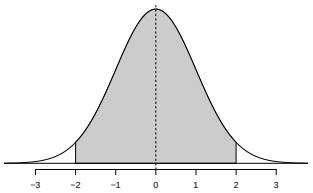
Working with your neighbor(s) see if you can work out the following probabilities for a random variable $Z \sim N(0, 1)$.

- $P(Z < -1)$
- $P(Z > 2.22)$
- $P(-1.53 \leq Z \leq 2.75)$
- $P(0.75 \geq Z \text{ or } Z \geq 1.43)$

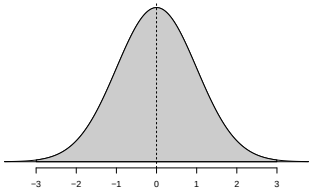
Empirical Rule



$P(-1 \leq z \leq 1) = \Phi(1) - \Phi(-1) = 0.683$



$P(-2 \leq z \leq 2) = \Phi(2) - \Phi(-2) = 0.954$



$P(-3 \leq z \leq 3) = \Phi(3) - \Phi(-3) = 0.997$

Standardizing Normal Distributions

Everything we just discussed applies only to the unit normal distribution, but this doesn't come up very often in problems.

Let X be a normally distributed random variable with mean μ and variance σ^2 then we define the random variable Z such that

$$Z = \left(\frac{X - \mu}{\sigma} \right) \sim N(0, 1)$$

In previous classes you have probably referred to this as a z-score. We will see why this works soon when we get to functions of random variables.

$$P(a \leq X \leq b) = P\left(\frac{a - \mu}{\sigma} \leq Z \leq \frac{b - \mu}{\sigma} \right) = \Phi\left(\frac{b - \mu}{\sigma} \right) - \Phi\left(\frac{a - \mu}{\sigma} \right)$$

Last time

We were able to show that for a binomial random variable as n gets large the probability mass function converges to

$$f(x) = \frac{1}{\sqrt{2\pi npq}} \; e^{-\frac{1}{2}\frac{(x-np)^2}{npq}}$$

which is the probability density function of a Normal distribution with $\mu = np$ and $\sigma^2 = npq$.

Therefore we can approximate probability calculations for a Binomial random variable using Normal probability calculations.

Back to the Aluminum case example

A company is testing a new manufacturing process for aluminum cases, if 20% of the cases do not meet their specifications what is the probability that if the company manufactured 1,000 cases at least 850 will be within specification?

Let X be the number of cases within spec out of the 1,000 then

$$P(X \geq 850) = \sum_{k=850}^{1000} \binom{1000}{k} (0.8)^k (0.2)^{1000-k} = 0.0000264$$

We can approximate this probability with a normal distribution where $\mu = np = 800$ and $\sigma^2 = npq = 160$.

$$P(X \geq 850) \approx P\left(Z \geq \frac{850 - np}{\sqrt{npq}}\right) = P(Z \geq 3.95) = 1 - \Phi(3.95) = 0.000038$$

Back to the Aluminum case example - cont.

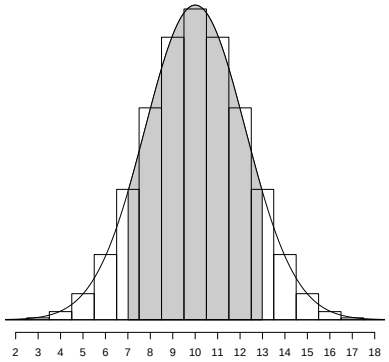
What if the company wants to know that at least 810 cases will be within specification?

$$P(X \geq 810) = \sum_{k=810}^{1000} \binom{1000}{k} (0.8)^k (0.2)^{1000-k}$$

$$P(X \geq 810) \approx P\left(Z \geq \frac{810 - np}{\sqrt{npq}}\right) = P(Z \geq 0.79) = 1 - \Phi(3.95) = 0.215$$

Improving the approximation

Take for example a Binomial distribution where $n = 20$ and $p = 0.5$, we should be able to approximate the distribution with $X \sim \mathcal{N}(\mu = 10, \sigma^2 = 5)$.



It is clear that our approximation is missing $1/2$ of $P(X = 7)$ and $P(X = 13)$, as $n \to \infty$ this error is very small. In this case $P(X = 7) = P(X = 13) = 0.073$ so our approximation is off by $\approx 7\%$.

Improving the approximation - cont.

Binomial probability:

$$P(7 \leq X \leq 13) = \sum_{x=7}^{13} \binom{20}{x} 0.5^x (1-0.5)^{20-x}$$

Naive approximation:

$$P(7 \leq X \leq 13) \approx \Phi\left(\frac{13-10}{\sqrt{5}}\right) - \Phi\left(\frac{7-10}{\sqrt{5}}\right)$$

Continuity corrected approximation:

$$P(7 \leq X \leq 13) \approx \Phi\left(\frac{13+1/2-10}{\sqrt{5}}\right) - \Phi\left(\frac{7-1/2-10}{\sqrt{5}}\right)$$

Improving the approximation - cont.

When we scale up n to 200 and our range interest to $P(70 \leq x \leq 130)$:
Binomial probability:

$$P(70 \leq X \leq 130) = \sum_{x=70}^{130} \binom{200}{x} 0.5^x (1-0.5)^{200-x}$$

Naive approximation:

$$P(70 \leq X \leq 130) \approx \Phi\left(\frac{130-100}{\sqrt{50}}\right) - \Phi\left(\frac{70-100}{\sqrt{50}}\right)$$

Continuity corrected approximation:

$$P(70 \leq X \leq 130) \approx \Phi\left(\frac{130+1/2-100}{\sqrt{500}}\right) - \Phi\left(\frac{70-1/2-100}{\sqrt{50}}\right)$$

Improving the approximation, cont.

This correction also lets us do moderately useless things like calculate the probability for a particular value of k. Such as, what is the chance of 50 Heads in 100 tosses of slightly unfair coin ($p = 0.55$)? Binomial probability:

$$P(X = 50) = \binom{100}{50} 0.55^{50} (1-0.55)^{50}$$

Naive approximation:

$$P(X = 50) \approx \Phi\left(\frac{50-55}{4.97}\right) - \Phi\left(\frac{50-55}{4.97}\right)$$

Continuity corrected approximation:

$$P(X = 50) \approx \Phi\left(\frac{50+1/2-55}{\sqrt{4.97}}\right) - \Phi\left(\frac{50-1/2-55}{\sqrt{4.97}}\right)$$

Sample Problem (1)

Roll a fair die 500 times, what's the probability of rolling at least 100 ones?

Sample Problem (2) - 2.2.13

A pollster wishes to know the percentage p of people in a population who intend to vote for a particular candidate. How large must a random sample with replacement be in order to be at least 95% sure that the sample percentage is within one percentage point of p?