The Normal Distribution

BIOS 6611

CU Anschutz

Week 3

Properties

Welcome to the Rest of Your Life!

Source: The Good Place

The Normal Distribution

The **normal distribution** (also called the *Gaussian distribution*) is one of the most important distributions in statistics

The normal distribution has unique properties that make it extraordinary:

- the central limit theorem
- the linear combination of normally distributed random variables is still normal
- it isn't terrible to work with mathematically

The Normal PDF

The PDF of the normal distribution is defined by two parameters:

- μ is the *mean* and represents the "location"
- σ^2 is the *variance* and represents the "squared scale", where $\sigma^2 > 0$

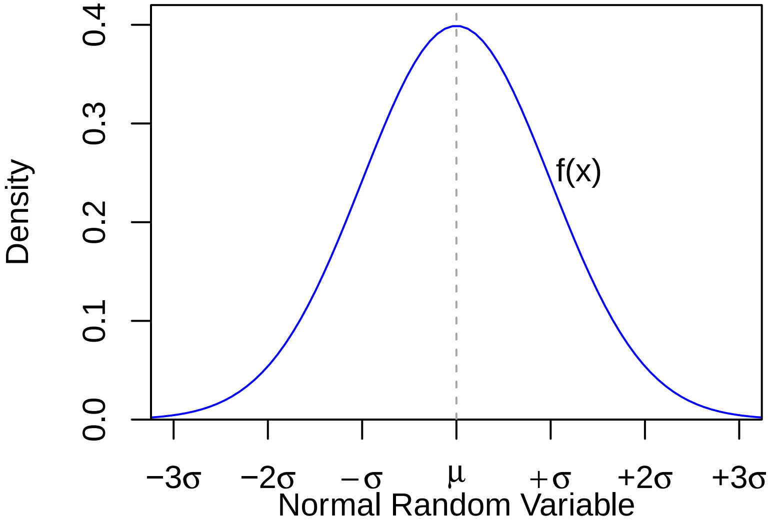
The *probability density function* is

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

From the PDF we could prove that $E(X) = \mu$ and $Var(X) = \sigma^2$ for the normal distribution

Normal Properties - I

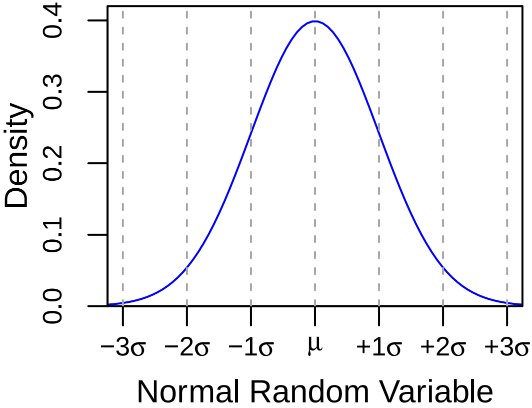
The normal distribution is symmetric about μ (i.e., $f(\mu + x) = f(\mu - x)$):



Normal Properties - II

This symmetry also helps with some quick rules of thumb for probability:

- $P(\mu - \sigma < X < \mu + \sigma) = 0.6827$ (about 68% lies between ± 1 s.d.)
- $P(\mu - 2\sigma < X < \mu + 2\sigma) = 0.9545$ (about 95% lies between ± 2 s.d.)
- $P(\mu - 3\sigma < X < \mu + 3\sigma) = 0.9973$ (about 99.7% lies between ± 3 s.d.)



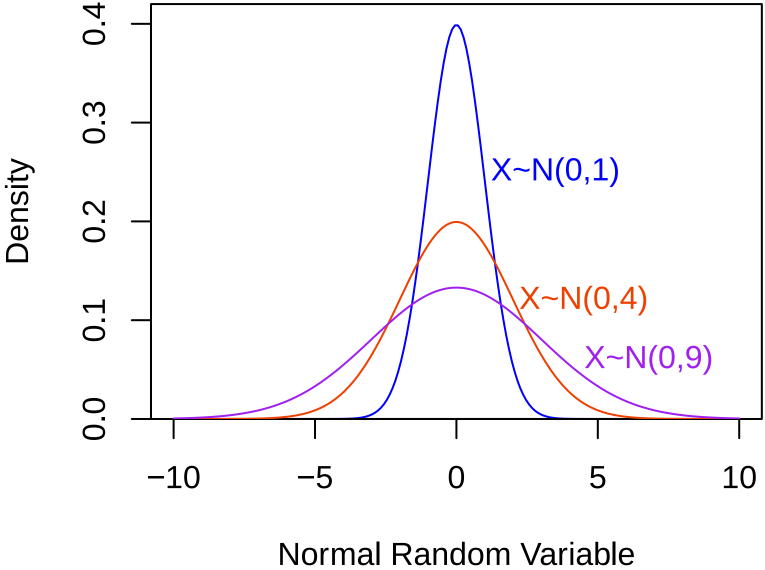
Normal Properties - III

The symmetry of the normal distribution also means that all three common summaries of central tendency (i.e., the mean, median, and mode) are all equal to μ

Quick Review:

- mean: the (arithmetic) average of our sample
- median: the middle value of our sample when arranged in order of magnitude (i.e., smallest to largest)
- mode: the most frequent value of our sample (i.e., the peak of the PDF)

The PDF Shape with Varying σ^2



Standard Normal Distribution

Standard Normal

The simplest form of the normal distribution is when $\mu = 0$ and $\sigma^2 = 1$: $X \sim N(0, 1)$. This is called the **standard normal distribution**.

The PDF for the standard normal distribution is

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

Here the PDF is simpler than the general form we saw earlier, where we now have $E(X) = 0$ and $Var(X) = 1$.

Standardizing a Normal R.V.

We can transform any normally distributed random variable (e.g., $X \sim N(\mu, \sigma^2)$) in a standard normal variable (e.g., $Z \sim N(0, 1)$):

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

Because the standard normal distribution is utilized so frequently, its PDF and CDF have special notation:

- $\phi(z)$ is our standard normal PDF
- $\Phi(z)$ is the standard normal CDF

$\Phi(z)$ of Yesteryear

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

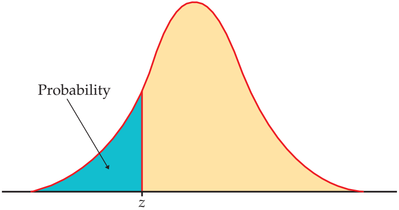


TABLE A

Standard Normal probabilities

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0003	.0002
−3.3	.0005	.0005	.0005	.0004	.0004	.0004	.0004	.0004	.0004	.0003
−3.2	.0007	.0007	.0006	.0006	.0006	.0006	.0006	.0005	.0005	.0005
−3.1	.0010	.0009	.0009	.0009	.0008	.0008	.0008	.0008	.0007	.0007
−3.0	.0013	.0013	.0013	.0012	.0012	.0011	.0011	.0011	.0010	.0010
−2.9	.0019	.0018	.0018	.0017	.0016	.0016	.0015	.0015	.0014	.0014
−2.8	.0026	.0025	.0024	.0023	.0023	.0022	.0021	.0021	.0020	.0019
−2.7	.0035	.0034	.0033	.0032	.0031	.0030	.0029	.0028	.0027	.0026
−2.6	.0047	.0045	.0044	.0043	.0041	.0040	.0039	.0038	.0037	.0036
−2.5	.0062	.0060	.0059	.0057	.0055	.0054	.0052	.0051	.0049	.0048
−2.4	.0082	.0080	.0078	.0075	.0073	.0071	.0069	.0068	.0066	.0064

$\Phi(z)$ (and more) of Today

Fortunately, we don't have to only rely on tables today. In R we can calculate the PDF ($\phi(z)$) with `dnorm()`, CDF ($\Phi(z)$) with `pnorm()`, or the value z is for a given area under the curve (AUC) with `qnorm()`:

```r
z <- -2.58
dnorm(z) #PDF
```

```
## [1] 0.01430511
```

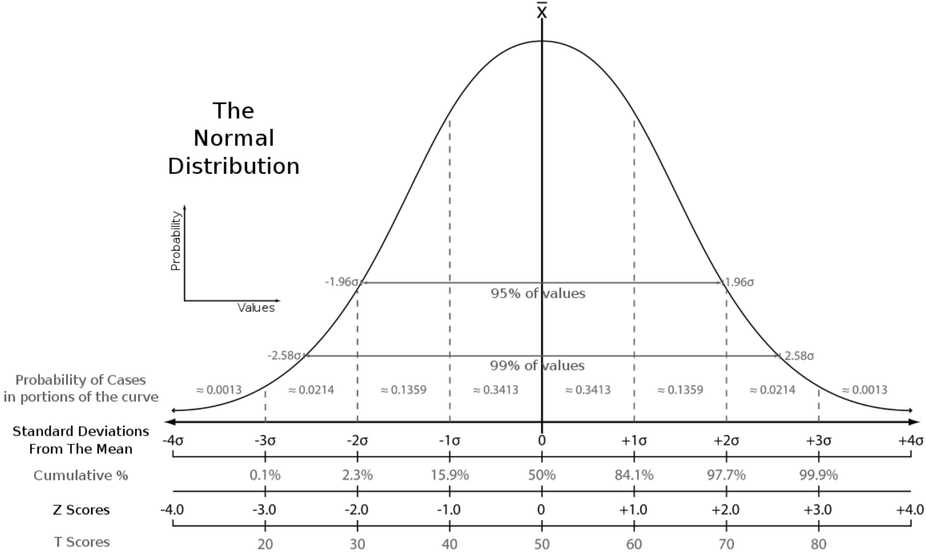
```r
pnorm(z) #CDF
```

```
## [1] 0.004940016
```

```r
qnorm(0.004940016) #quantile (i.e., z-score for a given AUC)
```

```
## [1] -2.58
```

Standard Normal PDF



Population Moments and Describing Distributions

Population Moments

So far we have discussed location and variability of distributions, but we can also describe the **skewness** and **kurtosis** of the population (or a sample also).

$E(X) = \mu$ 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 also called *central* moments).

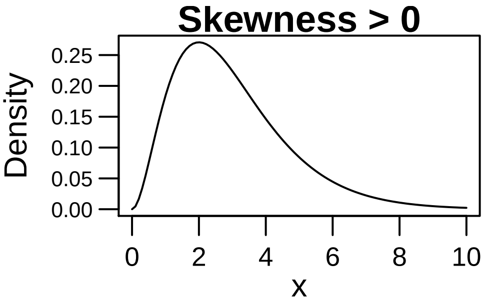
From the central moments we can represent the *standardized* moments by dividing the k^{th} standard deviation:

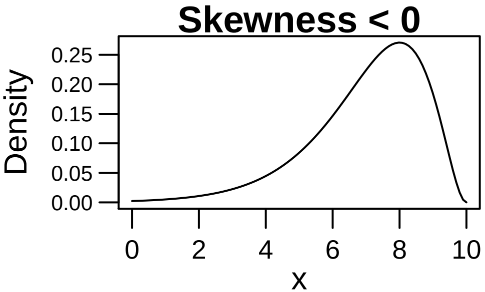
- $\frac{E\left[(X-\mu)^3\right]}{\sigma^3}$ represents the skewness
- $\frac{E\left[(X-\mu)^4\right]}{\sigma^4}$ represents the kurtosis

Skewness

Skewness describes the symmetry of the distribution.

- The 3rd central moment of the data, like the 1st, 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 is <0: negative skew; skewed to the left

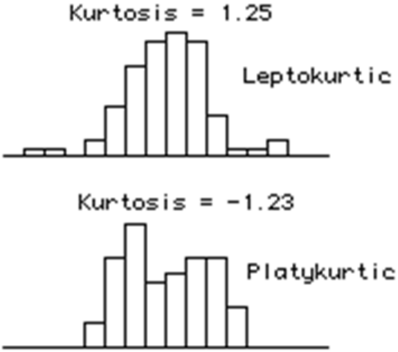




Kurtosis

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 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 Student's t, exponential, and Poisson distributions
- Excess Kurtosis < 0, lighter/thinner tails than a normal distribution (*platykurtic*); such as uniform and Bernoulli distributions



The More You Know (FYI)

Alternative Parameterization of the PDF

Some people prefer specifying the scale of then normal distribution in terms of its **precision** instead of the variance

The precision is the reciprocal of the variance: $\tau = \frac{1}{\sigma^2}$

The PDF in this parameterization is

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

One of the more common places you will see this parameterization is with Bayesian statistics