M3S2 - Normal Distribution

Professor Jarad Niemi

STAT 226 - Iowa State University

September 28, 2018

Outline

- Continuous random variables
  - normal
  - Student's t (later)
- Normal random variables
  - Expectation/mean
  - Variance/standard deviation
  - Standardizing (z-score)
  - Calculating probabilities (areas under the bell curve)
  - Empirical rule: 68%, 95%, 99.7%

Normal

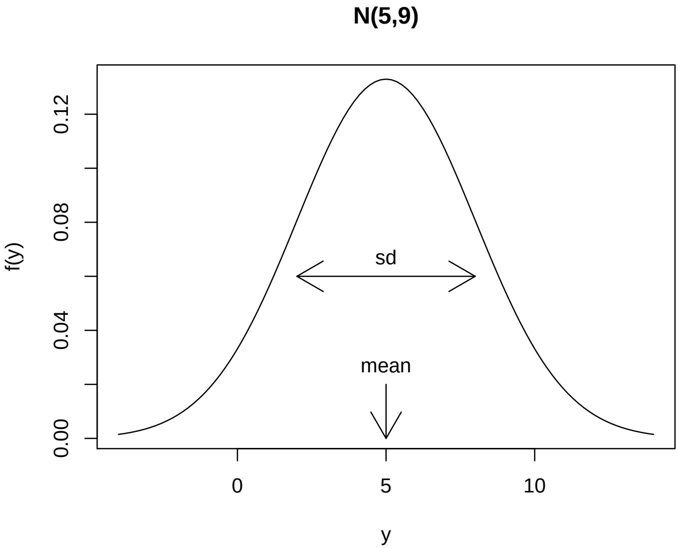
Definition

A normal random variable with mean μ and standard deviation σ has a probability distribution function

$$f(y) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{\left(-\frac{1}{2\sigma^2}(y-\mu)^2\right)}$$

for $\sigma > 0$ where $e \approx 2.718$ is Euler's number. A normal random variable has mean μ, i.e. $E[Y] = \mu$, and variance $Var[Y] = \sigma^2$ (and standard deviation $SD[Y] = \sigma$). We write $Y \sim N(\mu, \sigma^2)$.

Example normal pdf



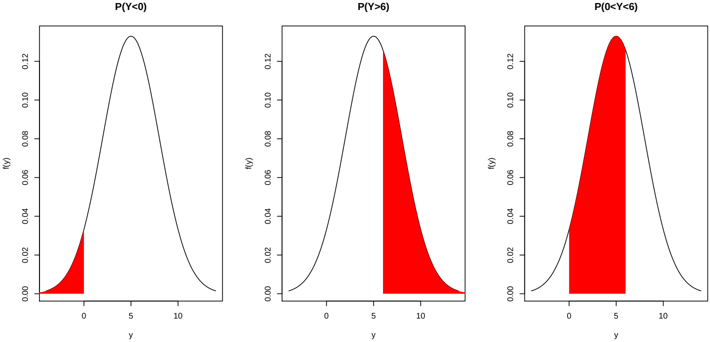
Interpreting PDFs for continuous random variables

For continuous random variables, we calculate areas under the curve to evaluate probability statements. Suppose $Y \sim N(5, 9)$, then

- $P(Y < 0)$ is the area under the curve to the left of 0,
- $P(Y > 6)$ is the area under the curve to the right of 6, and
- $P(0 < Y < 6)$ is the area under the curve between 0 and 6

where the curve refers to the bell curve centered at 5 and with a standard deviation of 3 (variance of 9) because $Y \sim N(5, 9)$.

Areas under the curve



Standardizing

Definition

A standard normal random variable has mean $\mu = 0$ and standard deviation $\sigma = 1$. You can standardize any normal random variable by subtracting its mean and dividing by its standard deviation. If $Y \sim N(\mu, \sigma^2)$, then

$$Z = \frac{Y - \mu}{\sigma} \sim N(0, 1).$$

For an observed normal random variable y, a z-score is obtained by standardizing, i.e.

$$z = \frac{y - \mu}{\sigma}.$$

z-tables exist to calculate areas under the curve (probabilities) for standard normal random variables.

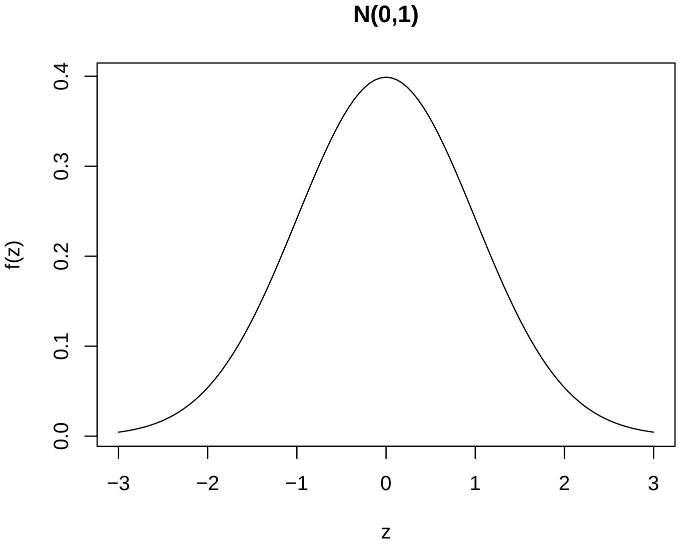

N(0,1)
f(z)
0.0
0.1
0.2
0.3
0.4
−3
−2
−1
0
1
2
3
z

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

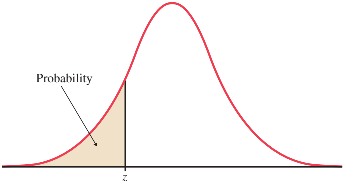


TABLE A Standard normal probabilities

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
−2.3	.0107	.0104	.0102	.0099	.0096	.0094	.0091	.0089	.0087	.0084
−2.2	.0139	.0136	.0132	.0129	.0125	.0122	.0119	.0116	.0113	.0110
−2.1	.0179	.0174	.0170	.0166	.0162	.0158	.0154	.0150	.0146	.0143
−2.0	.0228	.0222	.0217	.0212	.0207	.0202	.0197	.0192	.0188	.0183
−1.9	.0287	.0281	.0274	.0268	.0262	.0256	.0250	.0244	.0239	.0233
−1.8	.0359	.0351	.0344	.0336	.0329	.0322	.0314	.0307	.0301	.0294
−1.7	.0446	.0436	.0427	.0418	.0409	.0401	.0392	.0384	.0375	.0367
−1.6	.0548	.0537	.0526	.0516	.0505	.0495	.0485	.0475	.0465	.0455
−1.5	.0668	.0655	.0643	.0630	.0618	.0606	.0594	.0582	.0571	.0559
−1.4	.0808	.0793	.0778	.0764	.0749	.0735	.0721	.0708	.0694	.0681

Calculating probabilities by standardizing

Using z-tables, we can calculate the probabilities for any normal random variable.

Suppose $Y \sim N(\mu, \sigma^2)$ and we want to calculate $P(Y < c)$, then

$$P\left(Y < c\right) = P\left(\frac{Y-\mu}{\sigma} < \frac{c-\mu}{\sigma}\right) = P\left(Z < \frac{c-\mu}{\sigma}\right).$$

Since c, μ, and σ are all known, $\frac{c-\mu}{\sigma}$ is just a number.

In addition, we have the following rules

$$\begin{array}{lcl} P(Y > c) & = 1 - P(Y \le & c) \quad \text{probabilities sum to 1} \\ P(Y \le c) & = \qquad P(Y < & c) \quad \text{continuous random variable} \end{array}$$

Example z-table use

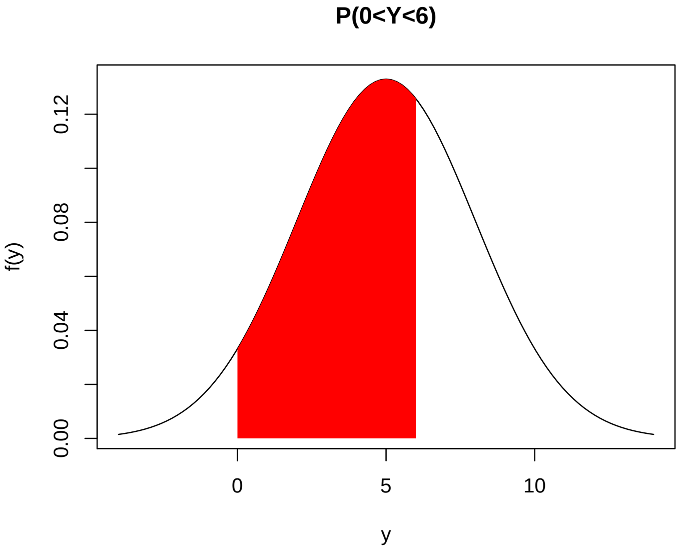
Suppose $Y \sim N(5, 9)$, then

$$\begin{aligned}
P(Y < 0) &= P\left(\frac{Y-5}{3} < \frac{0-5}{3}\right) && \text{standardize} \\
&\approx P\left(Z < -1.67\right) && \text{calculation} \\
&= 0.0475 && \text{z-table lookup}
\end{aligned}$$

$$\begin{aligned}
P(Y > 6) &= P\left(\frac{Y-5}{3} > \frac{6-5}{3}\right) && \text{standardize} \\
&\approx P\left(Z > 0.33\right) && \text{calculation} \\
&= 1 - P\left(Z < 0.33\right) && \text{probabilities sum to 1} \\
&= 0.3707 && \text{z-table lookup}
\end{aligned}$$

$$\begin{aligned}
P(0 < Y < 6) &= P(Y < 6) - P(Y < 0) && \\
&= [1 - P(Y > 6)] - P(Y < 0) && \text{probabilities sum to 1} \\
&= [1 - 0.3707] - 0.0475 && \text{previous results} \\
&= 0.5818 &&
\end{aligned}$$

Differences of probabilities



Inventory management

Suppose that based on past history Wheatsfield Coop knows that during any given month, the amount of wheat flour that is purchased follows a normal distribution with mean 20 lbs and standard deviation 4 lbs. Currently, Wheatsfield has 25 lbs of wheat flour in stock for this month. What is the probability Wheatsfield runs out of wheat flour this month?

Let Y be the amount of wheat flour purchased this month and assume $Y \sim N(20, 4^2)$. Then

$$\begin{aligned} P(Y > 25) &= P\left(\tfrac{Y-20}{4} > \tfrac{25-20}{4}\right) \\ &= P(Z > 1.25) \\ &= P(Z < -1.25) \\ &= 0.1056 \end{aligned}$$

There is approximately an 11% probability Wheatsfield will run out of wheat flour this month.

Empirical rule

Definition

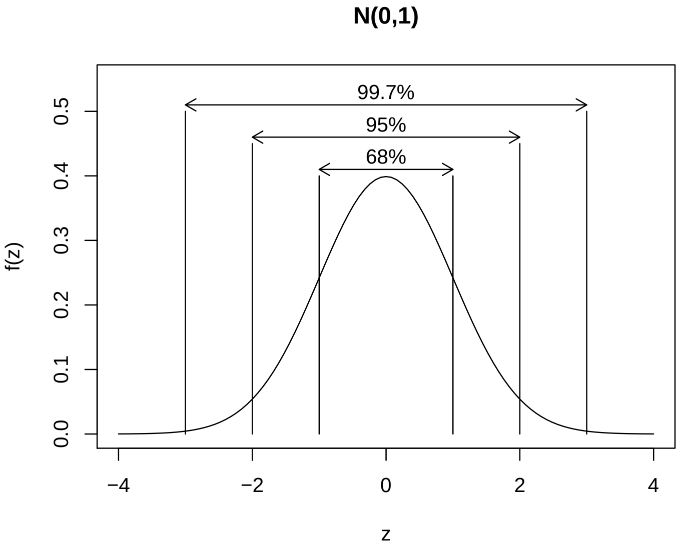
The empirical rule states that for a normal distribution, on average,

- 68% of observations will fall within 1 standard deviation of the mean,
- 95% of observations will fall within 2 standard deviations of the mean, and
- 99.7% of observations will fall within 3 standard deviations of the mean.

For a standard normal, i.e. $Z \sim N(0,1)$,

$$
\begin{aligned}
P(-1 < Z < 1) &= P(Z < 1) - P(Z < -1) \\
&= [1 - P(Z < -1)] - P(Z < -1) \\
&= 1 - 2 \cdot P(Z < -1) = 1 - 2 \cdot 0.1587 && \approx 0.68 \\
P(-2 < Z < 2) &= 1 - 2 \cdot P(Z < -2) = 1 - 2 \cdot 0.0228 && \approx 0.95 \\
P(-3 < Z < 3) &= 1 - 2 \cdot P(Z < -3) = 1 - 2 \cdot 0.0013 && \approx 0.997
\end{aligned}
$$

Empirical rule - graphically



Empirical rule

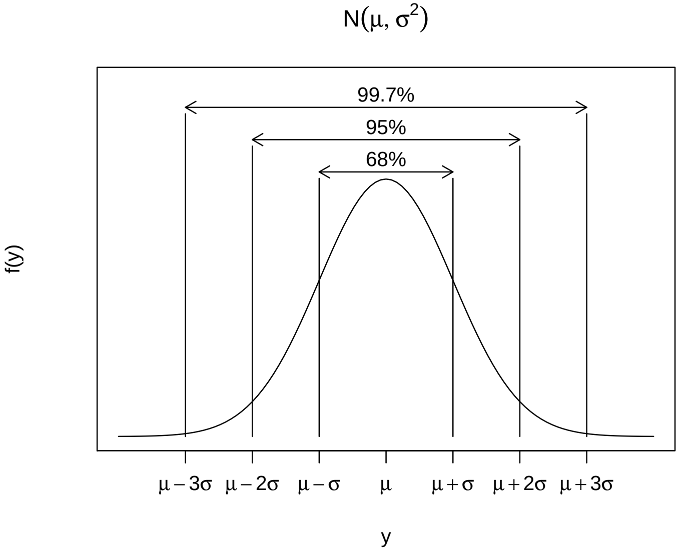
Let $Y \sim N(\mu, \sigma^2)$, then the probability Y is within c standard deviations of the mean is

$$P(\mu - c \cdot \sigma < Y < \mu + c \cdot \sigma) = P\left(-c < \frac{Y - \mu}{\sigma} < c\right) = P(-c < Z < c).$$

Thus

- 68% of observations will fall within 1 standard deviation of the mean,
- 95% of observations will fall within 2 standard deviations of the mean, and
- 99.7% of observations will fall within 3 standard deviations of the mean.

Empirical rule - graphically



CONAN

If we have two independent random normal variables $X \sim N(\mu_X, \sigma_X^2)$ and $Y \sim N(\mu_Y, \sigma_y^2)$, then

$$aX + bY + c \sim N(a\mu_X + b\mu_Y + c, a^2\sigma_X^2 + b^2\sigma_Y^2)$$

Thus, linear Combinations Of Normals Are Normal (CONAN). If you have a linear combination, all you need to do is find the expectation and variance of the linear combination using properties of expectations and variances, i.e.

$$\begin{aligned} E[aX + bY + c] &= a\mu_X + b\mu_Y + c \\ Var[aX + bY + c] &= a^2\sigma_X^2 + b^2\sigma_Y^2. \end{aligned}$$

We will use this later to find the sampling distribution of the sample mean when the underlying random variables are normally distributed.