

The Normal distribution

- The normal probability distribution is the most common model for relative frequencies of a quantitative variable.
- Bell-shaped and described by the function

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

where y is the quantitative variable whose relative frequencies we are modeling, μ is the mean of the distribution, and σ^2 is the variance.

- For fixed values of μ and σ (e.g., 0 and 1) we can evaluate $f(y)$ for a range of values of y . The plot of $f(y)$ against y has the familiar bell shape.
- Figures.

Normal distribution (cont'd)

- Examples of possibly normally distributed variables:
 - Sale price of houses in a given area
 - Interest rates offered by financial institutions in the US
 - Corn yields across the 99 counties in Iowa.
- Examples of variables that are *not* normally distributed:
 - Number of traffic accidents per week at each intersection in Ames.
 - Proportion of US population who voted for Bush, Gore, or Nader in 2000.
 - Proportion of consumers who prefer Coke or Pepsi.

Normal distribution (cont'd)

- If the normal distribution is a good model for the relative frequencies of y , then we say that “ y is distributed as a normal random variable” and write $y \sim N(\mu, \sigma^2)$. We will be interesting in estimating μ and σ^2 from sample data.
- Given estimates of μ and σ^2 , we can also estimate the *probability* that y is in the interval (a, b) .
- From calculus:

$$\text{Prob}(a < y < b) = \int_a^b f(y)dy.$$

- No need to know calculus! Tables of probabilities under the normal distribution can be used to calculate any probability of interest.
- Example: Table C.1 in text.
- Entries in the table give probability under the curve between 0 and z , where

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

- The variable z is called a *standard normal random variable* and its distribution is a normal distribution with mean $\mu = 0$ and variance $\sigma^2 = 1$.

Normal distribution (cont'd)

Example 1: $y \sim N(50, 225)$. Then

$$z = \frac{y - 50}{15} \sim N(0, 1).$$

- To compute the probability that y is between 35 and 70, we first "standardize" and then use the table:

1. Standardize by subtracting mean and dividing by the standard deviation. For $y = 70$:

$$z = \frac{70 - 50}{15} = 1.33,$$

and for $y = 35$:

$$z = \frac{35 - 50}{15} = -1.00.$$

2. Get area under curve between 0 and 1.33 and between -1.00 and 0, and add them up: $0.4082 + 0.3413 = 0.7495$.
3. Interpretation in "English": if y is normal with mean 50 and standard deviation 15, we expect that the probability that y is between 35 and 70 is about 75%.

Normal distribution (cont'd)

- Example 2: $y \sim N(10, 64)$. Find the probability that y is bigger than 15 and also the probability that y is less than 7.

$$\begin{aligned}\text{Prob}(y > 15) &= \text{Prob}\left(\frac{y - 10}{8} > \frac{15 - 10}{8}\right) \\ &= \text{Prob}(z > 0.625) \\ &= 1 - \text{Prob}(z < 0.625) \\ &= 1 - (0.5 + 0.2357) = 0.264.\end{aligned}$$

See figure.

$$\begin{aligned}\text{Prob}(y < 7) &= \text{Prob}\left(\frac{y - 10}{8} < \frac{7 - 10}{8}\right) \\ &= \text{Prob}(z < -0.375) \\ &= 1 - \text{Prob}(z > 0.375) \\ &= 1 - (0.5 + 0.1480) = 0.352.\end{aligned}$$

See figure.

Normal distribution (cont'd)

For any value of (μ, σ^2) ,

- y is equally likely to be above or below the mean:

$$\text{Prob}(y < \mu) = \text{Prob}(y > \mu) = 0.5),$$

because the normal distribution is symmetric about the mean.

- Because of symmetry, the mean, median and mode of a normal variable are the same.
- A normal random variable can take on any value on the real line, so that

$$\text{Prob}(-\infty < y < \infty) = 1.$$

- The probability that y is within a standard deviation of the mean is approximately 0.68:

$$\text{Prob}(-\sigma < y < \sigma) = \text{Prob}(-1 < z < 1) \approx 0.68.$$

and also:

$$\text{Prob}(-2\sigma < y < 2\sigma) = \text{Prob}(-2 < z < 2) \approx 0.95$$

and

$$\text{Prob}(-3\sigma < y < 3\sigma) = \text{Prob}(-3 < z < 3) \approx 0.99.$$

Sampling distributions

- We use sample data to make inferences about populations. In particular, we compute sample statistics and use them as estimators of population parameters.
- The sample mean \bar{y} is a ‘good’ estimator of the population mean μ and the sample variance S^2 (or $\hat{\sigma}^2$) is a ‘good’ estimator of the population variance σ^2 .
- How reliable an estimator is a sample statistic?
- To answer the question, we need to know the *sampling distribution* of the statistic.
- This is one of the most difficult concepts in statistics!

Sampling distributions (cont'd)

- Suppose that $y \sim N(\mu, 25)$ and we wish to estimate μ . We proceed as follows:
 1. Draw a sample of size n from the population: y_1, y_2, \dots, y_n .
 2. Compute the sample mean: $\bar{y} = n^{-1} \sum_i y_i$.
- Two things to note:
 1. The sample is *random*! If I had collected more than one sample of size n from the same population, I would have obtained different values of y . Then, \bar{y} is also a random variable.
 2. The larger n , the more reliable is \bar{y} as an estimator of μ .
- Example using simulation: pretend that we have $y \sim N(20, 25)$ and draw 30 random samples, each of size $n = 10$ from the population using the computer. With each sample, compute \bar{y} .

Sampling distributions (cont'd)

- The sampling distribution of a statistic computed from a sample of size $n > 1$ has smaller variance than the distribution of the variable itself.
- **Theorem:** If y_1, y_2, \dots, y_n are a random sample from some population, then:
 - Mean of \bar{y} equals mean of y : $E(\bar{y}) = \mu_{\bar{y}} = \mu$.
 - Variance of \bar{y} is smaller than variance of y : $\sigma_{\bar{y}}^2 = \frac{\sigma^2}{n}$. The larger the sample, the smaller the variance (or the higher the reliability) of \bar{y} .
- **Central Limit Theorem:** For large n , the sample mean \bar{y} has a distribution that is approximately normal, with mean μ and variance σ^2/n , *regardless* of the shape of the distribution from which we sample the y 's.
- The larger the sample, the better the approximation (see Figure 1.14).
- Example in lab.

Sampling distributions (cont'd)

- Given the sampling distribution of \bar{y} , we can make probability statements such as:
 - $\text{Prob}(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{y} < \mu + 2\frac{\sigma}{\sqrt{n}}) = 0.95$.
 - $\text{Prob}(\bar{y} > a) = \text{Prob}(z > \frac{a-\mu}{\sigma/\sqrt{n}})$ and use the standard normal table to get an answer.
- A preview of coming attractions: if it is true that

$$\text{Prob}(\mu - 2\frac{\sigma}{\sqrt{n}} < \bar{y} < \mu + 2\frac{\sigma}{\sqrt{n}}) = 0.95$$

then we can estimate the following interval using sample data:

$$(\bar{y} - 2\frac{\hat{\sigma}}{\sqrt{n}}, \bar{y} + 2\frac{\hat{\sigma}}{\sqrt{n}})$$

We call it a **95% confidence interval for the population mean**.

- As in the case of \bar{y} , the interval is also *random* and varies from sample to sample.
- If we drew 100 samples of size n from some population and computed 100 intervals like the one above, about 95 of them would cover the *true but unknown* value of μ .