

Continuous Probability Distributions

Recall that we have contrasted continuous variables from discrete variables by noting that while there are gaps between possible values of a discrete variable, a continuous variable can take any value in some interval of the number line.

Suppose X is a **discrete** random variable taking on the values $1, 2, \dots, n$ with **probability distribution function** $f_X(x)$. Then

$$P[2 \leq X \leq 4] = f_X(2) + f_X(3) + f_X(4)$$

Suppose X is a **continuous** random variable with **probability density function** $f_X(x)$. Then

$$P[a < X < b] = \int_a^b f(x) dx \quad (1)$$

A random variable X is said to follow a Gaussian (Normal) distribution with population parameters μ and σ if its probability density function is given by

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

A random variable X is said to follow the standard Normal distribution if its population parameters μ and σ equal 0 and 1, respectively. In this case,

$$f_X(x) = \frac{1}{\sqrt{2\pi}} e^{-\frac{x^2}{2}}$$

Suppose X is a continuous random variable taking on any value $x \in [c, d]$. Then the probability density function satisfies

$$\begin{aligned} f_X(x) &> 0 \quad \text{if } x \in [c, d] \\ &= 0 \quad \text{otherwise} \end{aligned}$$

$$\int_c^d f_X(u) du = 1,$$

The probability $P(a < X < b)$ equals

$$\int_a^b f_X(u) du$$

Example (Normal distribution)

Suppose X represents blood pressure and suppose that the population mean $\mu = 129$ and the population standard deviation $\sigma = 19.8$.

$$\begin{aligned} P[X > 150] &= \int_{150}^{\infty} \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}} \\ &= \int_{150}^{\infty} \frac{1}{\sqrt{2\pi(19.8)^2}} e^{-\frac{(x-(129))^2}{2(19.8)^2}} \end{aligned}$$

The probability that a Normal random variable takes on a value within an interval is equal to the area under the part of the normal density which lies above the interval.

Unfortunately, there is no simple formula for calculating this area, so we need to use a table.

Fortunately, we only need a table for the **standard normal density** with mean $\mu = 0$ and standard deviation $\sigma = 1$.

If we want to calculate probabilities for a general normal random variable X with mean μ and SD σ , we need to construct a new random variable, called the **standardized score** of X ,

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

Given values for μ and σ , we can actually go back and forth between the “ X scale” and the “ Z scale:”

$$X = \mu + \sigma Z.$$

Two simple rules can be very helpful in calculating normal probabilities:

- Since the total area under any density is 1,

$$P[Z > z] = 1 - P[Z \leq z].$$

- Since the normal density is symmetric about 0,

$$P[Z < z] = P[Z > -z] = 1 - P[Z \leq -z];$$

Suppose that X is a random variable that represents height. For the population of 18 to 74 year old women, height is normally distributed with mean $\mu = 63.9$ inches and standard deviation $\sigma = 2.6$ inches.

If we randomly select a woman from this population, what's the probability that she is between 60 and 68 inches tall ?

Suppose serum cholesterol levels X for children in Wisconsin have mean 175mg/100ml and SD 30mg/100ml. Suppose we want to know the limits within which 95% of the population lies.

We know $P[Z > 1.96] = 0.025$ so that
 $P[-1.96 \leq Z \leq 1.96] = 0.95$.

What kinds of questions can you now answer ?

Use BP as an illustration. Suppose we know BP is normally distributed with a specific mean μ and variance σ^2 .

1. A person walks in and you record the BP. You can tell if this person has “normal” BP or is an outlier.
2. You can tell what proportion of people lie inside or outside a given range.
3. If, say, 20 men come in to the office on a given day and each has his BP taken, you can tell the probability that at least one (at most 2, at least 5, etc.) lie outside or inside some given range.

Among females in the United States between 18 and 74 years of age, diastolic blood pressure is normally distributed with mean $\mu = 77$ mmHG and standard deviation $\sigma = 11.6$ mmHg.

1. What is the probability that a randomly selected woman has a diastolic blood pressure less than 60 mmHg ?
2. What is the probability that a randomly selected woman has a diastolic blood pressure greater than 90 mmHg ?
3. What is the probability that among five women selected at random from the population, at least one will have a pressure outside the range 60 to 90 mmHg ?

To answer these questions, we’ve made some key assumptions:

We have “known” that the populations of interest are Normally distributed.

We have “known” the population mean, μ , and the population standard deviation, σ

Most of the time, we don’t know these things. So most often we collect a random, independent, sample from a population and **estimate** population parameters of interest.

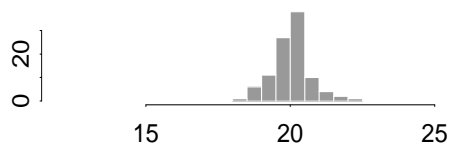
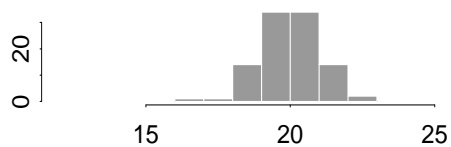
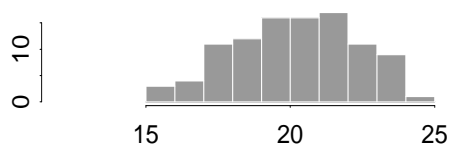
Statistical Inference:

The process of drawing conclusions about an entire population based on the information in a sample is known as **statistical inference**.

Q: I happen to “know” BP levels for all men in the U.S. follows a normal distribution with mean μ and standard deviation σ .

You need to guess at μ and σ . How would you do this ?

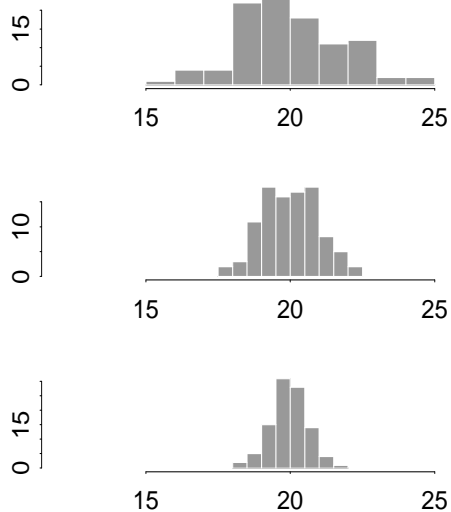
Histograms of 100 sample means for different sample sizes (n).



Q: I happen to know that the number of car accidents in Madison each year follows a Poisson distribution. I know the mean (and so I know the variance).

You need to guess at the mean. How would you do this ?

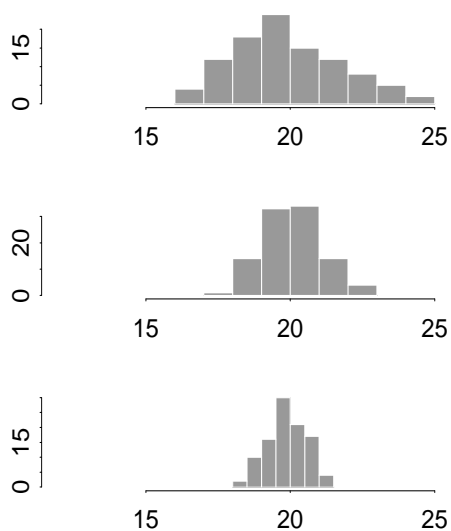
Histograms of 100 sample means for different sample sizes (n).



Q: I happen to know that the number of successful surgeries out of 10,000 follows a Binomial distribution. I know the mean and variance.

You need to guess at the mean and variance. How would you do this ?

Histograms of 100 sample means for different sample sizes (n).



The probability distribution of \bar{X} is called the **sampling distribution** of \bar{X} .

Understanding properties of the sampling distribution of \bar{X} allows us to make inference about population parameters based on a single sample!

Characteristics that we observed in histograms which approximate the sampling distribution of \bar{X}

1. The mean of the sampling distribution is near the population mean from which the samples were taken (sample size n doesn't matter).
2. The variance of the sampling distribution gets smaller as the size of the sample (n) increases.
3. For large sample sizes (n), the sampling distribution looks normal.

CENTRAL LIMIT THEOREM

Let X_1, X_2, \dots, X_n denote n independent random variables sampled from the same distribution which has a finite mean (μ) and variance (σ^2).

If n is large, then

$$\frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \approx Z \sim N(0, 1)$$

In other words, \bar{X} is approximately Normally distributed with mean μ and variance $\frac{\sigma^2}{n}$

Approximation gets better as n increases.

NOTE: n independent samples from the same distribution are often called independent and identically distributed (i.i.d.)

Recall the questions stated earlier on BP.

Is a given BP typical ?

What proportions of BPs lie in a given range ?

If you see 20 men in a day, what's the probability that at least one will have BP outside a given range ?

To answer these questions, we needed to assume that BP is normally distributed with a specific mean μ and variance σ^2 .

Thanks to the CLT, we can now answer similar questions without assuming a Normal distribution.

Q: Consider the distribution of cholesterol levels for US men aged 20 - 74. Assume the mean $\mu = 211$ mg/100ml and the standard deviation is $\sigma = 46$ mg/100ml. Select a sample of size n from the population.

Is the sample an outlier ? Does it have an unusually high or low sample mean ?

To answer this, we could figure out the interval that encloses say 95% of the sample means and see if the sample mean from our sample falls within that range.

So for a fixed n , we want to find x_l and x_u such that

$$P[x_l \leq \bar{X}_n \leq x_u] = 0.95. \text{ We know } P[-1.96 \leq Z \leq 1.96] = 0.95$$

For $n = 25$,

$$\begin{aligned} P[x_l \leq \bar{X} \leq x_u] &= P[-1.96 \leq \frac{\bar{X} - 211}{46/\sqrt{25}} \leq 1.96] \\ &= P[-1.96 \cdot \left(\frac{46}{\sqrt{25}}\right) \leq \bar{X} - 211 \leq 1.96 \cdot \left(\frac{46}{\sqrt{25}}\right)] \\ &= P[211 - 1.96 \cdot 9.2 \leq \bar{X} \leq 211 + 1.96 \cdot 9.2] \\ &= P[192.97 \leq \bar{X} \leq 229.03] \end{aligned}$$