PSY 503: Foundations of Psychological Methods

Lecture 12: Estimation and Uncertainty II

Robin Gomila

Princeton

October 3, 2020

Central Limit Theorem

The distribution of the sample mean approaches the normal distribution as the sample size increases.

Central Limit Theorem

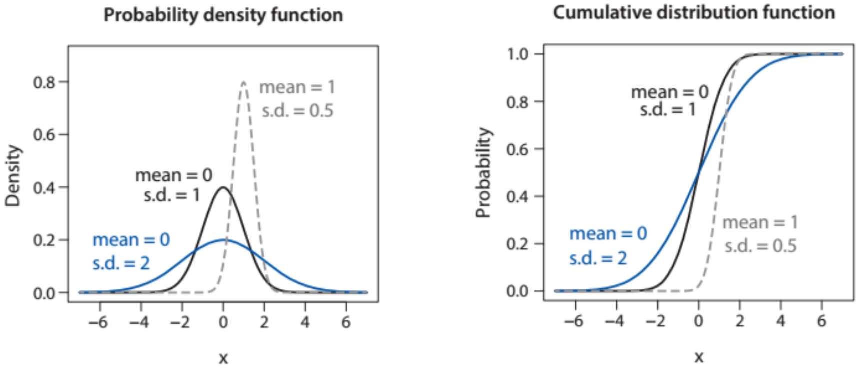
- Kicks in regardless of the distribution of the random variable
- This result is incredibly useful because the normal distribution is a parametric distribution
- This allows us to quantify the uncertainty of our estimates!

The normal distribution

- Also called the *Gaussian distribution*
- Can take any number on the real line $(-\infty, \infty)$
  - Continuous distribution
- Two parameters:
  - mean μ
  - standard deviation σ
- If X is a random variable, we may write

$$X \sim N(\mu, \sigma^2)$$

The normal distribution: PDF and CDF



- PDF: bell shaped, centered and symmetric around μ
- Standard deviation "controls" for the spread of the distribution
- Different means shift the PDF and CDF without changing their shape
- Larger standard deviations mean more variability

FYI: PDF of the normal distribution

$$f(x;(\mu,\sigma)) = \frac{1}{\sigma\sqrt{2\pi}}\exp\Big(-\frac{1}{2}(\frac{x-\mu}{\sigma})^2\Big), \quad -\infty \leq x \leq \infty$$

Standard normal distribution

The *standard normal distribution* is a normal distribution with $\mu = 0$ and $\sigma = 1$

The normal distribution: Properties

1. Adding a constant to (substracting a constant from) a normal random variable yields another normal random variable
2. Multiplying / dividing a random variable by a constant yields another normal random variable

The normal distribution: Properties

Let $X \sim N(\mu, \sigma^2)$. Let c be a constant. Then the following properties hold:

1. A random variable defined by $Z = X + c$ also follows a normal distribution, with $Z \sim N(\mu + c, \sigma^2)$
2. A random variable defined by $Z = cX$ also follows a normal distribution, with $Z \sim N(c\mu, (c\sigma)^2)$

Implication

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

is normally distributed.

This is the formula for the z-score of X, which represents the number of standard deviations an observation is above vs. below the mean.

As a result,

$$\text{z-score} = \frac{X - \mu}{\sigma} \sim N(0, 1)$$

The normal distribution: Properties

- If a random variable is defined on $X \sim \mathbb{N}(\mu, \sigma^2)$ (independently of the values of μ and σ):
  - the area under the normal curve between $\mu - \sigma$ and $\mu + \sigma$ is about 0.68. Formally, $P(\mu - \sigma \leq X \leq \mu + \sigma) = 0.68$
  - the area under the normal curve between $\mu - 1.64\sigma$ and $\mu + 1.64\sigma$ is about 0.90. Formally, $P(\mu - 1.64\sigma \leq X \leq \mu + 1.64\sigma) = 0.90$
  - the area under the normal curve between $\mu - 1/96\sigma$ and $\mu + 1.96\sigma$ is about 0.95. Formally, $P(\mu - 1.96\sigma \leq X \leq \mu + 1.96\sigma) = 0.95$
  - the area under the normal curve between $\mu - 2.58\sigma$ and $\mu + 2.58\sigma$ is about 0.99. Formally, $P(\mu - 2.58\sigma \leq X \leq \mu + 2.58\sigma) = 0.99$
  - the area under the normal curve between $\mu - 3\sigma$ and $\mu + 3\sigma$ is about 0.997. Formally, $P(\mu - 3\sigma \leq X \leq \mu + 3\sigma) = 0.997$

Central Limit Theorem

Suppose that we obtain a random sample of n i.i.d. observations $X_1, X_2, ..., X_n$ from a probability distribution with mean $\mathbb{E}[X]$ and variance $\mathbb{V}[X]$.

Let's denote the sample average $\bar{X_n} = \sum_{i=1}^{n} \frac{X_i}{n}$.

Then the **central limit theorem** states that the sample mean converges in distribution to the normal distribution. We write

$$\bar{X_n} \to N(\mathbb{E}[X], \frac{\mathbb{V}[X]}{n})$$

Confidence intervals

- Range of values that are likely to include the true value of the parameter
- Researcher needs to decide the confidence level
  - Degree to which they'd like to be certain that the interval actually contains the true value of the parameter
  - Over a **hypothetically repeated data-generating process**, CIs contain the true value of the parameter with the probability of the confidence level (e.g., 95% confidence level)
- Confidence level often written $(1-\alpha) * 100\%$, where α can take any value between 0 and 1
  - e.g., $\alpha = .05$ corresponds to the 95% confidence level

Confidence intervals

- How do we calculate the 95% confidence of the sample mean for a sufficiently large sample of observations?
- Using the CLT, we know that sample mean is normally distributed
- Lower value of confidence interval is

$$[\bar{X}_n - 1.96 \times \mathsf{SD}(\bar{X})] = [\bar{X}_n - 1.96 \times \text{standard error}]$$

- Upper value of confidence interval is

$$[\bar{X}_n + 1.96 \times \mathsf{SD}(\bar{X})] = [\bar{X}_n + 1.96 \times \text{standard error}]$$

- 95 % confidence Interval is

$$[\bar{X}_n - 1.96 \times \text{standard error}, \bar{X}_n + 1.96 \times \text{standard error}]$$

Confidence intervals

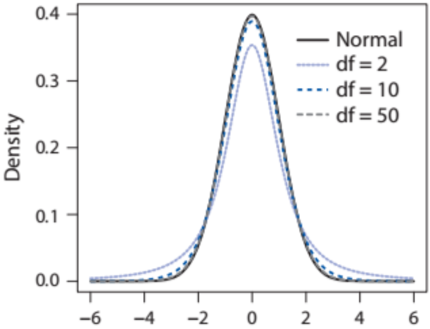
Let's open R Studio...

Confidence intervals and p-values

- Two sides of the same coin
- Suppose a random sample, a sample mean, and a null hypothesis that the sample mean is different from k. We will reject the null hypothesis at the significance level α (i.e., $p < \alpha$) if and only if the confidence interval with confidence level $1 - \alpha$ does not include k
- This applies to difference-in-means estimator
  - e.g., We reject the null hypothesis that there is no difference between the treatment and control groups at the 95% confidence level if and only if the 95% CI of the difference-in-means does not include zero, which corresponds to a p-value for a hypothesis test lower than .05

Small samples: the t-distribution

- Wondering how all of this relate to the t-distribution or t-tests?



- Small samples: more conservative test
  - t-distribution has fatter tails
  - coverage is more conservative

Small samples: the t-distribution

In small samples, the sampling distribution of the mean of variable X follows a t-distribution if and only if X is normally distributed!!!

This implies that until your sample is large enough for the CLT to kick in, t-tests won't work unless you assume that X is normally distributed. Most often NOT the case!

Illustration: Difference-in-means estimator

Go to R Studio...