

Lecture 19: Statistics & data analysis

(*Recipes*, Chapters 14 and 15*)

- Two types of problem

Probability theory is a rigorous branch of mathematics

→ allows us to calculate the probability of observing t if θ is true: $P(t | \theta)$

Statistics is a much less rigorous branch of mathematics and deals to a large extent with the inverse problem: given a measured value of t , what can we say about θ ?

*see also, *Statistics: a Guide to the Use of Statistical Methods in the Physical Sciences*, by Roger J. Barlow (Wiley)

Statistics and data analysis

Example of a result in probability theory: if the average number of radioactive decays occurring during some time interval is μ , the probability of observing n such events is $P(n | \mu) = \mu^n e^{-\mu}/n!$

Important notes:

μ is a continuous parameter (can take any non-negative real value)

n is an observable (random variable) and is an integer

P is not symmetric under interchange of n and μ (except in the limit of large n and μ)

Classical versus Bayesian statistics

- In statistics, there are two approaches to answering “given a measured value of t , call it t_m , what can we say about θ ?”
- Classical statistics approach:
 - There is a fixed, unknown value of θ , and all we can do is to quote values of θ , call them θ_1 and θ_2 , for which we know the probabilities of finding $t \leq t_m$ and $t \geq t_m$ in many repeated experiments: $\theta_1 < \theta < \theta_2$ is called the confidence interval

Classical versus Bayesian statistics

- In statistics, there are two approaches to answering “given a measured value of t , call it t_m , what can we say about θ ?”
- Bayesian statistics approach:
 - θ is a random variable that can be described by a probability distribution. Our experimental results and some ASSUMPTIONS allow us to describe the distribution

Classical versus Bayesian statistics

Bayesian and classical statistics meet in 2 places

1) Systems described by a Gaussian probability distribution

$$P(t | \theta) = \exp(-[t-\theta]^2/2\sigma^2) / (2\pi)^{1/2}\sigma = P(\theta | t)$$

i.e. P is symmetric in t and θ

→ Classicists integrate over t

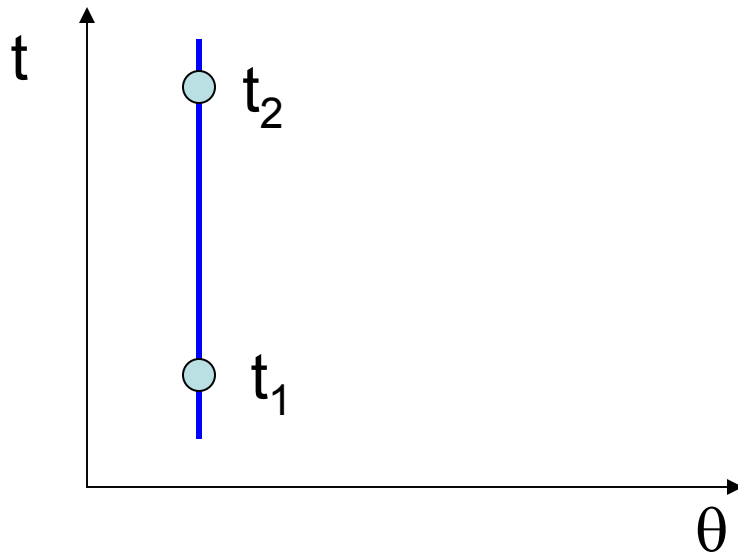
Bayesians integrate over θ

Classical versus Bayesian statistics

Bayesian and classical statistics meet in 2 places

2) Properly constructed 1-D problems

Measurements define confidence intervals which can be determined by the Neyman construction



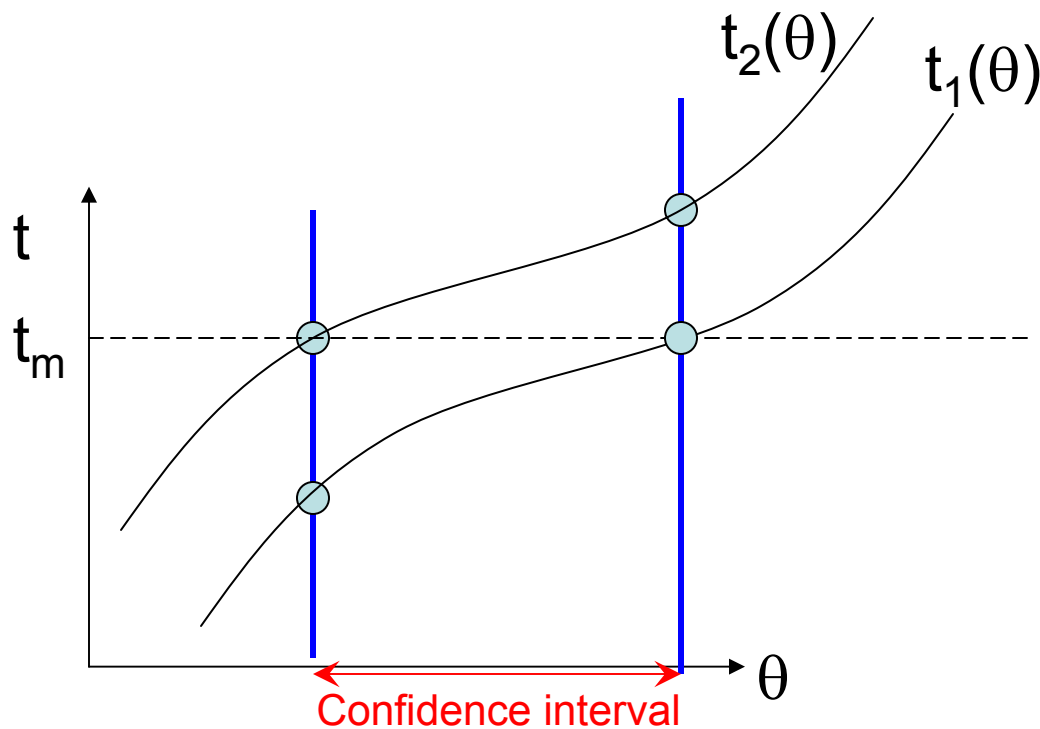
Make many vertical lines and mark points t_1 and t_2 such that

$$P(t > t_2 \mid \theta) = \alpha$$

$$P(t < t_1 \mid \theta) = \alpha$$

Confidential intervals for 1-D problems

- Neyman construction: $\int_{t_2}^{\infty} P(t | \theta) dt = \int_{-\infty}^{t_1} P(t | \theta) dt = \alpha$



Confidential intervals for 1-D problems

- This construction defines a central confidence interval of probability content $\beta = 1 - 2\alpha$
- A common convention is to choose $\beta = 0.683$ ($\pm 1\sigma$ for a Gaussian)

Confidential intervals for 1-D problems

- The center of the confidence interval is not uniquely defined: common choices are

a) A symmetrized interval, $\theta = \bar{\theta} \pm \Delta\theta$, where
 $\bar{\theta} = (\theta_1 + \theta_2) / 2$ and $\Delta\theta = |\theta_1 - \theta_2| / 2$

b) Use the 50% value defined by $\int_{t_{50}}^{\infty} P(t | \theta) dt = \int_{-\infty}^{t_{50}} P(t | \theta) dt = 0.5$

to write $\theta = \theta_{50 - (\theta_{50} - \theta_1)}$ ^{$+(\theta_2 - \theta_{50})$}

Bayesian statistics

- The Neyman construction is also something that a Bayesian can love:

Bayesian statistics starts with

$$P(A \cap B) = P(A|B) P(B) = P(B|A) P(A)$$

(Bayes' theorem)

In our language, we can write this

$$P(t|\theta) dt \cdot P(\theta) d\theta = P(\theta|t) d\theta \cdot P(t) dt$$

Bayesian statistics

- Our definition of t_1 and t_2 implies

$$\int_{t_1}^{t_2} P(t | \theta) dt = 1 - 2\alpha = \beta$$

- Multiply by $P(\theta)$ and integrate $d\theta$ to obtain

$$\int_{-\infty}^{\infty} d\theta P(\theta) \int_{t_1}^{t_2} dt P(t | \theta) = \beta \int_{-\infty}^{\infty} P(\theta) d\theta$$

$$\Rightarrow \int_{-\infty}^{\infty} d\theta \int_{t_1(\theta)}^{t_2(\theta)} dt P(t | \theta) P(\theta) = \beta$$

Bayesian statistics

- Apply Bayes' Theorem, and interchange order of integration:

$$\int_{-\infty}^{\infty} dt P(t) \int_{\theta_1(t)}^{\theta_2(t)} d\theta P(\theta | t) = \beta$$

- Suppose we measure the value $t = t_m$
Then $P(t) = \delta(t - t_m)$, and we can write

$$\int_{\theta_1(t_m)}^{\theta_2(t_m)} P(\theta | t_m) d\theta = \beta$$

Bayesian statistics

- This equation $\int_{\theta_1(t_m)}^{\theta_2(t_m)} P(\theta | t_m) d\theta = \beta$

says that given the measurement, $t = t_m$, the random variable θ has a probability β of lying between $\theta_1(t_m)$ and $\theta_2(t_m)$

Holds for integrated confidence intervals in 1-D

Summary

- By understanding our experiment, we know $P(t|\theta) dt$
- We would like to know $P(\theta|t_m) d\theta$ for a given measurement $t=t_m$
- We can compute

$$\int_{\theta_1(t_m)}^{\theta_2(t_m)} P(\theta | t_m) d\theta = \beta$$

“Bad” Bayesian statistics

- Classic fallacy is to assume that
$$P(\theta|t) = P(t|\theta) P(\theta) / P(t) = P(t|\theta)$$

Example: we observe n radioactive decays within a given time period Δt and assume

$$P(\mu|n) = P(n|\mu) = n^\mu e^{-n/\mu}$$

to obtain confidence limits on the decay rate $\mu/\Delta t$

Basic definitions

- Our textbook devotes several sections to describing the comparison of measured distribution functions of random variables. The basic definitions are useful:

- Mean $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$

- Variance, $Var(x) = \frac{1}{N-1} \sum_{i=1}^N (x_i - \bar{x})^2 \approx \langle x^2 \rangle - \langle x \rangle^2$

- Standard deviation, $\sigma = Var(x)^{1/2}$

Basic definitions

- Skew $Skew(x) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^3 / \sigma^3$
- Kurtosis $Kurt(x) = \frac{1}{N} \sum_{i=1}^N (x_i - \bar{x})^4 / \sigma^4 - 3$
- These 3rd and 4th moments are conventionally defined to be dimensionless and equal to zero for a Gaussian distribution

The book discusses the comparison of the means and variances of different distributions (Student's t-test and F test)

Basic definitions

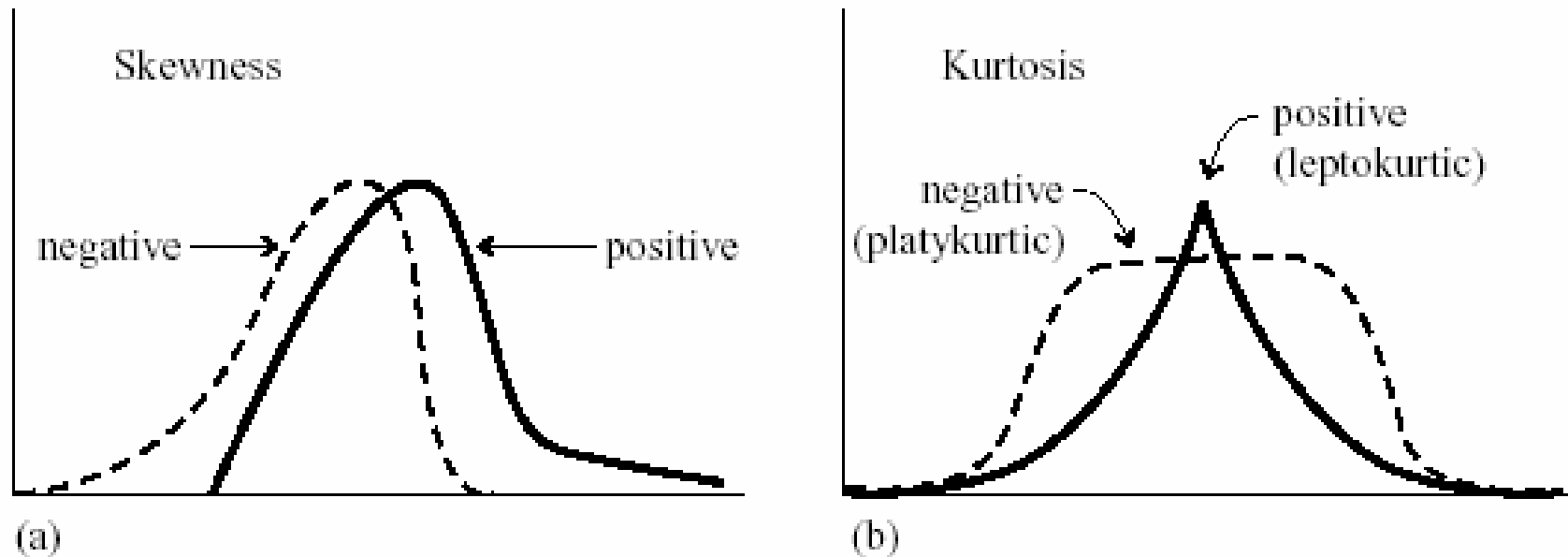


Figure 14.1.1. Distributions whose third and fourth moments are significantly different from a normal (Gaussian) distribution. (a) Skewness or third moment. (b) Kurtosis or fourth moment.

from *Recipes*

Central limit theorem

- An important property of the variance is that it is additive (like the mean)
- If $z = x + y$, where x and y are random variables drawn independently from different distributions, then

$$\bar{z} = \bar{x} + \bar{y}$$

$$\text{Var}(z) = \text{Var}(x) + \text{Var}(y)$$

- Hence, if X is the sum of N random variables x_i drawn from a set of arbitrary distributions, $R_i(x_i)$, then

$$\bar{X} = \sum_{i=1}^N \bar{x}_i \qquad \text{Var}(X) = \sum_{i=1}^N \text{Var}(x_i)$$

Central limit theorem

- The CLT says that in the limit of large N, the distribution of X is Gaussian:

$$P(X) = \frac{e^{-(X-\bar{X})^2/2\sigma^2}}{\sqrt{2\pi\sigma^2}}$$

- This justifies many of the Gaussian approximations commonly made in statistics

Central limit theorem

- Example: the mean of a sample $\bar{x} = \frac{1}{N} \sum_{i=1}^N x_i$

itself tends to be normally distributed with a “standard error”

$$\delta\bar{x} = \sqrt{\frac{\sigma_x^2}{N}} = \sqrt{\frac{\text{Var}(x)}{N}}$$

Central limit theorem: “proof”

- Consider a variable x_1 that has a probability distribution $P_1(x_1)$, and a variable x_2 that has a probability distribution $P_2(x_2)$

Let $X = x_1 + x_2$

$$\text{Then } P(X) dX = \int P_1(x_1) P_2(X - x_1) dx_1 dX$$

$$\rightarrow P = P_1 * P_2 \rightarrow \tilde{P} = \tilde{P}_1 \tilde{P}_2$$

(where tilde denotes a Fourier transform)

Central limit theorem: “proof”

- We know that

$$\begin{aligned}\tilde{P}_1(k) &= \int e^{ikx_1} P_1(x_1) dx_1 \\ &= \int P_1(x_1) dx_1 + ik \int x P_1(x_1) dx_1 - \frac{1}{2} k^2 \int x^2 P_1(x_1) dx_1 + \dots \\ &= 1 + ik\bar{x}_1 - \frac{1}{2} k^2 \langle x_1^2 \rangle + O(k^3)\end{aligned}$$

Central limit theorem: “proof”

- Taking the natural log, and using

$\ln(1+y) = y - \frac{1}{2}y^2 + O(y^3)$, we find that

$$\begin{aligned}\ln \tilde{P}_1(k) &= ik \langle x_1 \rangle - \frac{1}{2}k^2 \langle x_1^2 \rangle + \frac{1}{2}k^2 \langle x_1 \rangle^2 + O(k^3) \\ &= ik\bar{x}_1 - \frac{1}{2}k^2 \text{Var}(x_1) + O(k^3)\end{aligned}$$

- Thus the probability distribution for $X=x_1 + x_2$ has

$$\begin{aligned}\ln \tilde{P}(k) &= \ln \tilde{P}_1(k) \tilde{P}_2(k) = \ln \tilde{P}_1(k) + \ln \tilde{P}_2(k) \\ &= ik(\bar{x}_1 + \bar{x}_2) - \frac{1}{2}k^2 \{ \text{Var}(x_1) + \text{Var}(x_2) \} + O(k^3) \\ &= ik\bar{X} - \frac{1}{2}k^2 \text{Var}(X) + O(k^3)\end{aligned}$$

Central limit theorem: “proof”

- This expression is clearly true when X is the sum of any number of random variables

$$\ln \tilde{P}(k) = ik\bar{X} - \frac{1}{2}k^2 \text{Var}(X) + O(k^3)$$

Key point: as we add more and more random variables, the probability distribution gets broader and broader → its Fourier transform is described better and better by the terms of lowest order in k

Central limit theorem: “proof”

- So in the limit of large N, we take

$$\ln \tilde{P}(k) = ik\bar{X} - \frac{1}{2}k^2 \text{Var}(X)$$

$$\Rightarrow \tilde{P}(k) = e^{ik\bar{X}} e^{-\frac{1}{2}k^2 \text{Var}(X)}$$

Central limit theorem: “proof”

- The inverse Fourier transform yields

$$\begin{aligned}\tilde{P} &= \frac{1}{2\pi} \int_{-\infty}^{\infty} e^{-ikx} e^{ik\bar{X}} e^{-\frac{1}{2}k^2\sigma^2} dk \\ &= \frac{1}{2\pi} \int_{-\infty}^{\infty} \exp(-\frac{1}{2}[k\sigma - i(\bar{X} - x)/\sigma]^2) e^{-\frac{1}{2}[(\bar{X} - x)/\sigma]^2} dk \\ &= \frac{1}{2\pi} \sqrt{(2\pi/\sigma^2)} e^{-\frac{1}{2}[(\bar{X} - x)/\sigma]^2} \\ &= \frac{e^{-\frac{1}{2}[(\bar{X} - x)/\sigma]^2}}{\sqrt{2\pi\sigma^2}}\end{aligned}$$